

A visual exploration of the online conversational space

TOWARDS AN INTERACTIVE VISUALIZATION TOOL TO SUPPORT
JOURNALISTS' ANALYSIS OF USER GENERATED CONTENT.

PIETRO STEFANO LODI

POLITECNICO DI MILANO



FACOLTÀ DEL DESIGN

Pietro Stefano Lodi

Matr. 797056

M. Sc. in Communication Design

A. Y. 2013 / 2014

28 April 2015

Supervisor: Paolo Ciuccarelli

Co-Supervisor: Liliana Bounegru

ABSTRACT

Since the last decade journalism has been going through a fast and substantial revolution. The morning paper is now replaced with almost real-time news on the Internet. In this digital revolution, social media play an important role since they created a space where citizens have the same opportunities (broad audience) as reporters. These social networking sites can be seen as conversational spaces, where interactions between users and shared content reflect not only opinions, but eye witnessed material as well. It became pretty clear in the last few years that journalism should make use of social networking not only to increase the traffic to the website, but as a source of information to improve reporting.

However, these platforms continuously produce an information overload that makes it impossible for journalist to dive into. What they are missing is the clarity of bigger picture, which is necessary in order to see those elements that need further investigation.

This work's goal is therefore to design a tool able to support journalists and academics in both the exploration and investigation of such "ambient information stream". The project aims to help in two directions: on the one hand by providing an overview and on the other hand by giving the possibility to reach a granular level of information. The first one will be fulfilled by using a visual layer, following de Rosnay's notion of *macroscope*. The second one will be actualized by a structured level of interaction, able to show details-on-demand.

The combination of the two finds in Information Visualization the technique able to carry out those journalism's needs that this project aims to answer.

Nell'arco dell'ultimo decennio il giornalismo ha dovuto affrontare un cambiamento rapido e importante. Il quotidiano della mattina è ora sostituito da notizie online disponibili quasi in tempo reale. In questa rivoluzione digitale i social media svolgono un ruolo importante in quanto creatori di uno spazio dove i cittadini (utenti) hanno le stesse opportunità (il potenziale di rivolgersi alla massa) a cui fino a prima solo i giornalisti avevano accesso. Le piattaforme su cui i social media esistono possono essere viste come spazi conversazionali, dove le interazioni tra utenti e contenuti condivisi riflettono non solo opinioni, ma divengono anche fonte di materiale postato da testimoni oculari.

Negli ultimi anni si è mostrato con forte evidenza che il giornalismo dovrebbe cominciare a fare uso dei social media non solo per aumentare il numero di visite al sito web, ma in particolare come fonte di informazioni per migliorare l'attività di reporting.

Lo scopo di questo lavoro è dunque di progettare uno strumento di supporto a giornalisti e ricercatori sia nell'esplorazione che nell'investigazione di questo "flusso di informazioni". Il progetto mira ad aiutare in due direzioni: in primo luogo fornendo una visione complessiva e in secondo luogo fornendo la possibilità di ottenere un livello di informazione specifica. Il primo verrà aggiornato attraverso l'uso di un piano visivo, tenendo a mente la nozione di 'macroscopio' di de Rosnay. Il secondo invece verrà realizzato tramite l'utilizzo di una struttura interattiva, seguendo la tecnica del 'details-on-demand'.

La combinazione dei due trova nell'Information Visualization l'approccio in grado di realizzare quei bisogni del giornalismo che questo lavoro mira a rispondere.

Table of Contents

I ANALYSIS:

THE EVOLUTION OF JOURNALISM AND ITS NEW NEEDS

11 Digital journalism

13 From paper to digital

18 Consuming news

28 Prosuming news?

31 Journalists and social media

33 An introduction to Social Media

38 Social media in the news system

46 News sources

55 The needs

57 The target: breaking news reporters, analytic journalists, and academics

61 Research methods: investigation and verification

74 Interviews

86 Opportunities

89 Visualization: a methodology

91 A visual layer: simplify, explore and investigate

97 Visualizing Information

105 The role of visualization in the knowledge process

111 Interactivity

II SYNTHESIS: A SUPPORTIVE TOOL IN THE EXPLORATION OF ONLINE CONVERSATIONAL SPACES

119 **Understanding**

- 121 Brief
- 125 First thoughts
- 132 Existing tools and case studies

141 **Process**

- 143 Data Transformations: from raw data to data tables
- 152 The Visual Mapping loop
- 166 Visual Structures: Network
- 173 Visual Structures: Activity Map
- 177 Visual Structures: Content Timeline
- 181 View Transformations: Interface

189 **Final design**

205 **Conclusion**

- 207 Validation
- 211 Looking back to look forward

214 **Bibliography**

The evolution
of journalism
and its
new needs

Digital Journalism

FROM PAPER TO DIGITAL

The journalist profession did not face particular hard moments since its birth: newspapers were the way people had to get news, to stay informed and to be in touch with what was happening in the world.

Journalists and news organizations were those who had access to sources of information impossible to reach for an average person and were those who had the power to decide what to write in their article, therefore what citizens would have known about a specific story. This was the power newspapers had: whatever was printed on the paper meant to be true and trustful.

Peculiar in pre-digital forms of journalism (print and broadcast) was a strong separation between the author of the article and its audience. The communication was definitely uni-directional: from one (news organization) to many (the audience), top-down. Once an article was printed on the paper or broadcasted by radio or TV, there was no space for the reader/listener/spectator(s) to share his comments or opinion. Mail correspondence was the only option and was anyway reaching the author only, leaving this back-communication completely hidden from the public. Though every technology improvement made information richer and easier and faster to share, they nonetheless relied on a relatively small number of people to produce content and send it on its way.

From its earliest days, the Internet clearly appeared as a fundamentally different communication system. It was created with its core in the principle of democracy, to be decentralized, with power and control diffuse. Explains Tim Berners-Lee, inventor of the World Wide Web:

“The primary design principle underlying the Web’s usefulness and growth is universality. When you make a link, you can link to anything. That means people must be able to put anything on the Web, no matter what computer they have, software they use or human language they speak and regardless of whether they have a wired or wireless Internet connection. [...] Decentralization is another im-

portant design feature. You do not have to get approval from any central authority to add a page or make a link. All you have to do is use three simple, standard protocols: write a page in the HTML (hypertext markup language) format, name it with the URL naming convention, and serve it up on the Internet using HTTP (hypertext transfer protocol). Decentralization has made widespread innovation possible and will continue to do so in the future.”¹

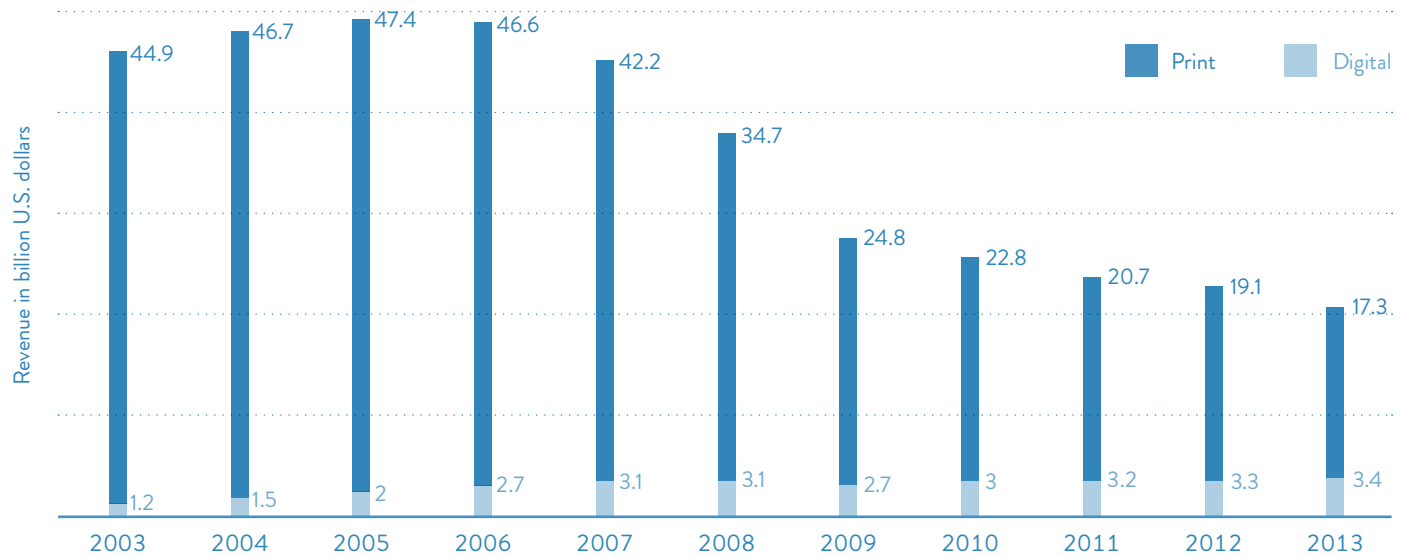
With the raise of the web as a communication medium, news organizations started to use it for publishing their articles. Newspapers were having a new life on the Internet and this brought many advantages. Says Steven Waldman:

“Lower barriers to entry and the vast amount of available space online have led to a greater diversity of voices, increased depth of some types of coverage, more consumer choices. Technology has reduced the costs of gathering, producing and distributing news, in some cases substantially. Reporters can use computerized databases to pull together stories in hours that would have previously taken weeks. The cost of producing and publishing images, sound, and text has fallen sharply. And most obviously, and most dramatically, the search-engine-driven Internet has made it infinitely easier to find a wide range of information rapidly.”

The Internet, however, brought together many new challenges. As a democratic medium, the Internet not only made publishing cheaper, but advertising too. The economical advantages, together with brand new features (unlike with print ad, advertisers were now able to know how many people were seeing and actually clicking on the banner) led to a huge shift of advertisements from the print to the web. On the web newspapers gained new audience and a flood of new ad dollars, but those were unable to make up for the loss in profits from the print products. The

1. T. Berners-Lee. 2010. *Long Live the Web: A Call for Continued Open Standards and Neutrality*. Scientific American. <http://goo.gl/jNLWNL>

Newspapers Association of America says that between 2005 and 2010 the Internet seriously cut newspaper revenue: even though the online ad revenue for the entire newspaper industry grew by a billion, print advertising lost \$24.6 billion (data from the *Newspapers Association of America, Advertising Expenditures*). The result was that “print dollars were being replaced by digital dimes.”



[1] Digital and Print advertising revenue of U.S. newspapers from 2003 to 2013. Data from *Statista.com*.

Next to the economical issues, which I will not investigate in this work, publishing on the web gave birth to new metrics useful for the newsroom to get insights on their work. Unlike the print edition (where number of sold units and locations they were sold in were the only available data), the Internet made possible to count the number of times each article is being read, from which location and the amount of time a user spends on the website or on the article’s page.

Although the web gave to newsrooms a more precise and instant way to know their public better, the author-audience relation didn’t change. News organizations saw the web as the new medium to publish with, without considering carefully its greater potential. This could be related due to three main reasons: economics, status and technology. The first one relies on the data discussed previously, where it

clearly appears that print editions pay more than the web ones. The second reason is the status. As the web became more popular, new space for people to publish their stories was created. Bloggers were taking it and they were spreading average citizens ideas, independent from any organization. Traditional journalists started to see bloggers and citizen journalists as inferior, worthless and even dangerous. They started to mark up their differences and to refer to the status print informa-

“Two characteristics of online news as opposed to traditional news are interactivity and immediacy.”

– Michael Karlsson

tion had over the Internet one. It also important not to forget the third reason: new technologies have always been welcomed with certain reluctance. New things ask people to change their habits, to change the way they have been working for a long time and many of us simply do not like to adapt to changes, always ready to point out the differences and difficulties. Hence it does sound perfectly clear why newsrooms took time to adapt to the new information changes.

When journalism goes online, observed Mark Deuze², it shares aspects of hypertextuality, multimediality and interactivity, changing and broadening its basic nature.

In the early 2000s a new step into the democracy of web was made: the web 2.0, also known as social media, started its diffusion. This big innovation changed the Internet landscape and its audience. To be more precise, the Internet became both author and audience, making the separation line between them more and more blurred. “Millions of people became not only consumers of information, but

2. M. Deuze. 2003. *The web and its journalisms: considering the consequences of different types of newsmedia online*. *New Media & Society* 5:2, 203-230. SAGE Publications.

creators, curators, and distributors.”³ Instead of information being provided primarily by a few large players, the ecosystem now could support millions of smaller players each serving a small but targeted audience. The era of platforms for self-expression and inter communication had begun, “a digital culture of public participation, re-mixing by individuals of data and information, harnessing the power of collective intelligence and providing services, rather than products.”⁴

Social media created new needs that the society of the web firmly embraces. Users now want to share their opinions and comments publicly, they want their voice to reach authors and other users. This whole change shifted the discussions from offline (those that take place in a real environment, with people facing each other, within physical walls) to online, a conversation anybody could join and to which anybody can add his opinion to or reply to specific comments by others.

The time for personal communication in the ‘letters to the author’ form has passed. It’s now the moment to recognize that a story does not end with its publication, but it’s rather a starting point for generating comments and contributions by the public. As Grabowicz points out: “For news organizations, Web 2.0 means moving away from using the Internet to draw a passive audience to a static publishing platform, and instead embracing the broader network, where communication, collaboration, interaction and user-created content are paramount.”⁵

-
3. S. Waldman. 2011. *Information Needs of Communities: The Changing Media Landscape in a Broadband Age*. DIANE Publishing.
 4. P. Grabowicz. 2014. *The Transition To Digital Journalism*. The Knight Digital Media Center, University of California.
 5. Ibid.

CONSUMING NEWS

“Fully describing the current media landscape is impossible; failing to try is irresponsible”, says Steven Waldman at the beginning of his “The changing media landscape in a broadband age” for *The information needs of communities*. And so do I at the beginning of this work, where I must acknowledge the whole media landscape is a very intricate picture, continuously changing in a fast progressing environment. New products and services, functions and features come out every day, creating new opportunities and new ways to interact with the information. It is basically open to users’ creativity and custom habits. Therefore this section aims to portrait the media landscape from afar, hoping to give a sufficient context to frame my work.

As the information distribution and publication changed, so did users behavior on news consumption. “The migration of news and information to an online platform has disrupted old patterns of reading and changed the relationship between audiences and news providers.”⁶ The Internet revolution not only moved the information from paper to digital screens, but it modified the way we make use of it. Let’s take for example the transitions journalism faced in the past. With radio and TV the information stopped being a personal moment, making it more social. Families and groups of people started to get together around those devices that were spreading information. Still nowadays the TV News often represents a family moment, a time in the everyday life for ‘watching the news’. The Internet brought back to a personal dimension of this moment. Personal, but not private. A personal moment because the information passes through the screens of our own devices: personal computers, smartphones, and tablets. Devices with which we interact personally in moments that are unlikely to be shared with others (the same goes

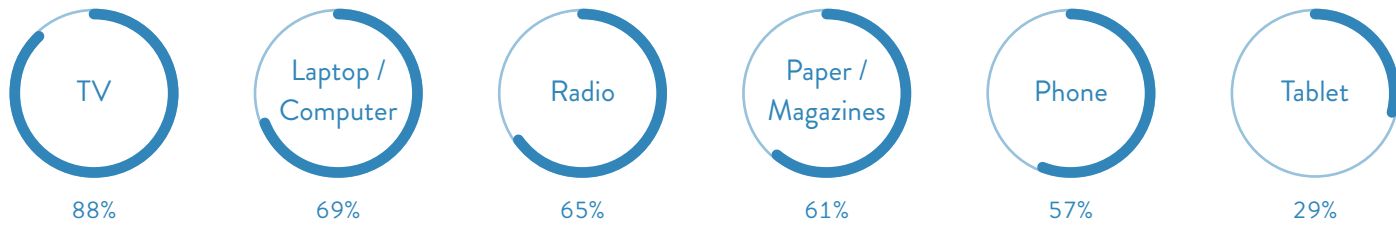
6. S. D. Reese; L. Rutigliano; K. Hyun; J. Jeong. 2007. *Mapping the blogosphere: Professional and citizen-based media in the global news arena*. Journalism 8:3, 235-261. SAGE Publications.



[2] Picture from: BBC News and James Harding, 2015. *The future of news.* <http://goo.gl/KaASYX>.



[3] Picture from: BBC News and James Harding. 2015. *The future of news*. <http://goo.gl/KaASXY>.



[4] Percentage of Americans using each device to get news. Data from *The media insight project*. 2014. American Press Institute. <http://goo.gl/71adBg>

for the old news-paper). However this new way we consume news is far from being private: users now comment under news articles, comment on Facebook shared links or put likes. They tweet their opinion with the Twitter community. It became part of users daily routine to share the articles they have read with a public (of close chosen friends or as big as the social media community), bringing the news consumption to an open, publicly shared dimension.

At the same time it changed also the pattern of when in the day people get their news. If with media like radio and TV it is the medium itself to define the rhythm of news consumption (there are specific appointments with the news during the day), the Internet brings it to a personal level, different for each user. This is possible due to a substantial change in the news publication: the web lets news organization (and any user) to publish content at any time of the day or night. Is the greatest thing that could happen for breaking news: information can start to flow as soon as the first author press the 'publish' button, no need to wait for the next TV News appointment anymore. News and information are continuously produced by millions of users and news organization, creating a huge amount of data that circulates on the web. Eric Schmidt, former CEO of Google, estimated that humans now create as much information in two days as we did from the appearance of Homo sapiens through 2003⁷.

Because information creation got fragmented, so did news consumption. Not anymore few defined moments in the day, the same ones for everybody, but many more, as many and lasting as much as each user wants. Upon this, the web lets everybody

7. M.G. Siegler (quoting Eric Schmidt). *Tech Crunch*. Aug 4, 2010. <http://goo.gl/VRirBG>

look for stories of their interest: the searching option that is now at the base of most digital interfaces (consider in example the importance of search engines) represents a revolution to the way people consumed information. Even though newspapers homepages are designed similarly to their paper first page, and sections are interactive links instead of indicating a page number, articles' content is search-

“It’s hard to find a new technology that news organisations don’t embrace. Read the Los Angeles Times on kindle. Watch the ABC News on YouTube. Leave a comment on a blog. [...] Listen to a podcast of ‘On Science’ from National Public Radio. Participate in a discussion board hosted by the Washington Post about college admissions. Receive SMS news about the Dallas Cowboys from The Dallas Morning News. Get features from Time on a PDA and tweets of breaking news from CNN.”

– Bob Franklin

able within an archive. Information is not anymore (not per se at least) based upon a prior choice on the newspaper name, but more frequently is searched on Google where different options from different newspapers are given.

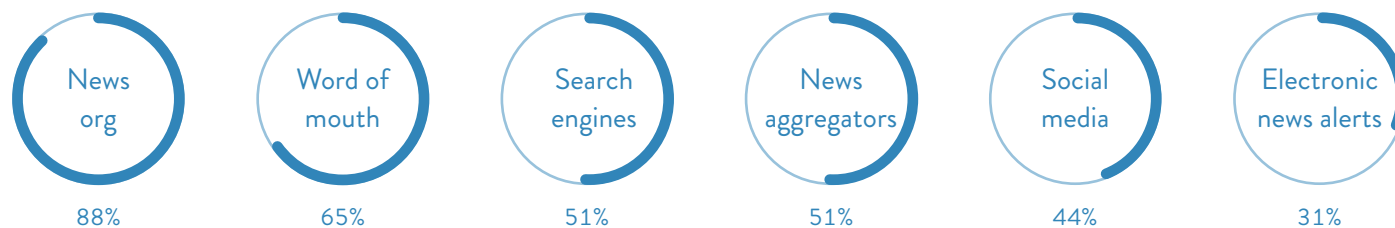
We now consume information through different devices, as a 2014 research conducted by the Media Insight Project shows. The telephone survey was conducted in the USA with the participation of 1492 adults (18+) from January 9 through February 16, 2014. Acknowledging that data from this project are representative of only a portion of the American citizens and are not wide enough to build solid conclusions, I personally think they are a good starting point (and certainly the

most updated one at this moment) to look at users behavior when it comes to media consumption. As reported by *the American Press Institute*, which together with *the Associated Press-NORC Center for Public Affairs Research* started the project, the survey was designed to probe what adults distinguish most in their news consumption in the digital age. These data “offer a portrait of Americans becoming increasingly comfortable using technology in ways that take advantage of the strengths of each medium and each device.”

It is important to highlight how wide is the number of devices utilized by users for getting news: where TV still represents the most widely used, within the top 6 we see computers, radio, print press (newspapers or magazines), cellphones and tablets. And the only major drop happens when it comes to tablets, which are used for news almost half as cellphones. A glance at the chart shows that most users make use of multiple devices for consuming news. We must also consider that each of those devices is also used in different situations (smartphones are a constant company throughout the day, whereas TV for example concerns much more a domestic context) and that the survey was conducted among a wide range of random users (different by age, race, politic preferences, income, education, employment situation). Hence we have an overview of the average American citizen and how it consumes news. We might venture that this survey pictures closely enough the average news consumer. It is important to keep this in mind when thinking about the content that flows online.

The news consumption panorama is not yet fully described until we start considering the way the information is presented to the user. It is not just about different devices, now every user can decides which is the way of news consumption that suits him/her at best.

There are two basically two ways in which users get in touch with information, and I would argue that these do not change since the birth of journalism. There are users who look for news and to stay up-to-date with what happens ‘out there’



[5] Percentage of Americans discovering news with each method. Data from *The media insight project*. 2014. American Press Institute. <http://goo.gl/71adBg>

and those instead who consume news when they are reached by them. We could say there are two main and opposite behaviors: active and passive. Active users are those who buy the paper every morning, who wait for the TV-news appointment or who check a news website on the Internet daily (and or throughout the day).

Those inactive instead receive information and news only when they become relevant and they get them from friends or because they became a sort of ‘topic of discussion’. This attitude is the one that many teenagers of these days have in front of news. Many literature points at this phenomenon as the proof that the next generation do not really care about journalism or news, while Waldman brilliantly highlights:

“Most news websites are free; friends can send links to you with a click of the mouse; news headlines appear before our eyes, unsolicited, on portals like Yahoo or AOL; free news apps on mobile devices find and display news from around the Internet. It should come as no surprise, then, when young people these days say they do not feel the need to seek out news sources, because if something important happens ‘the news will find me’.”⁸

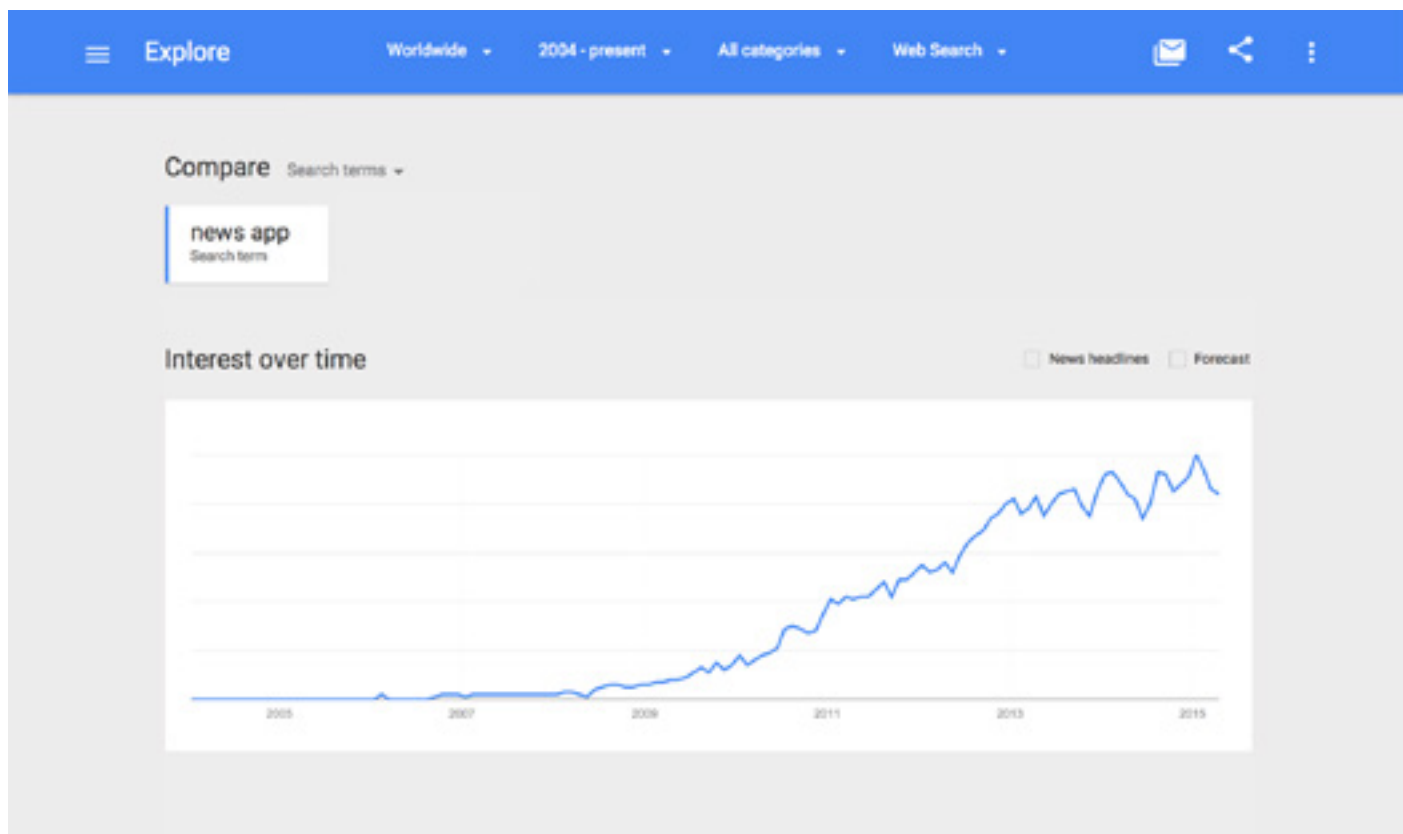
Of course this is not the most active attitude when thinking of news consumption, but it is a quite common behavior and a sign of how the Internet has changed the news world.

8. S. Waldman. 2011. *Information Needs of Communities: The Changing Media Landscape in a Broadband Age*. DIANE Publishing.

Talking about the active behavior, the news industry moved a lot in this direction. Serving the user when he wants to get informed and helping him/her in getting the news from the favorite sources is clearly a big business. News websites develop mobile-friendly versions of their websites to meet the needs of users while on-the-go, news aggregators continuously thrive on the idea that a better information comes from the combination of different news organizations' content (like the well known *GoogleNews*). And then we have apps. Yes, also the news industry entered in the growing market of mobile apps. After all, the opportunities given by these technologies are enormous: users can read their news in a more comfortable way (apps are usually better designed compared to the website, both from a UI, UX and a readability point of view), they can get real time notifications for breaking news or for whatever they express to be of their interest.

[6] Interest over time for the search term 'news app' on Google. *Google Trends*. Visited last in April 2015.

In some cases it turned out to be the new solution to make people subscribe to



Flipboard, which defines itself as: “your personal magazine, with the world’s best sources organized into thousands of topics, it’s a single place to follow the stories and people that matter to you.”⁹ With *Flipboard* users have the possibility to follow a specific news story and get more articles about that in the future, to keep informed without the necessity of looking for more. The app, after your approval, connects to your social networks profiles and peeks at what your circle of contacts shares. In this way it ensures to give you meaningful (or at least related to you and your friends) stories to browse through.

Another example is *Snapchat*. This photo messaging app has nothing to do with news. Until the beginning of 2015, when on January 27 the company released *Discovery*, a new built-in feature that lets users enjoy content from news websites. Its goal is not to provide breaking news articles, but to give a new and playful way to enjoy news content. “There’s something new everyday”¹⁰ says the voice of the launching spot. Their secret is “a new way to explore Stories from different editorial teams. It’s the result of collaboration with world-class leaders in media to build a storytelling format that puts the narrative first.”¹¹ Simplicity in news stories consumption is the key for *Snapchat Discovery*’s concept and yet another trace of how differently the audience is willing to experience information.

9. Flipboard website’s home page. Accessed on February 20th 2015. <https://flipboard.com/>

10. *Introducing Discover* advertisement published on YouTube. <http://goo.gl/VRle0H>

11. Snapchat Blog. January 27, 2015. *Introducing Discovery*. <http://goo.gl/kYUxw0>

PROSUMING NEWS?

In the 1980 book *The Third Wave*, Alvin Toffler uses for the first time the term ‘prosumer’. He defined the prosumer as someone who blurs the distinction between a ‘consumer’ and a ‘producer’. Alvin was clearly a futuristic mind, since we still use the term to express something that thanks to the Internet actually exists. In fact the web 2.0 revolution introduced the possibility to non-professional users to produce content available for the public. “The foundation of the participatory media revolution is that all people, not just those selected by an official system of some kind, have access to the means to publish and broadcast.”¹² This probably made the Internet the most wide and democratic space of all, where the principles outlined as the base of the World Wide Web, were finally actualized.

“Where traditionally journalism was charged with monitoring and reflecting public expression, citizens can now hold those conversations among themselves and, in a new twist, amplify the ‘conversations’ among journalists.”¹³

“When everyone can be a publisher,
what distinguishes the journalist?”

– Arthur S. Hayes

The question of these era is: ‘Who is producing news?’. Since the raise of bloggers and social media, the distinction between professionals and non-professionals, became more and more difficult to locate. We might say that those who are part of the professional journalism deal with “a combination of features including a claim to ‘authority’ and the command of economic resources available to media organi-

12. P. Miel; R. Faris. 2008. *News and information as digital media come of age*. Berkman Center for Internet & Society at Harvard University. Cambridge, USA.

13. S. D. Reese; L. Rutigliano; K. Hyun; J. Jeong. 2007. *Mapping the blogosphere: Professional and citizen-based media in the global news arena*. Journalism 8:3, 235-261. SAGE Publications.

zations.”¹⁴ In these terms appears a bit easier to make a distinction, even though the digital world we live in is mixing content and forms of communication leaving the question open to discussions. Journalists can now post news and articles from their own blog, in the same forms and on the same platform as common citizens do. I therefore ask to myself: is it necessary to mark the differences? Is it right to try to separate the Internet information in two, separated groups? As discussed before, the web put on the same level author and audience, giving to the latter the same tools and possibilities in the hands of the first. On social media like Twitter there is a mix of content level (I here refer to the meaningful level of each tweet) mainly due to the different nature of its users: news organizations, journalists, cultural organizations, politicians, celebrities, TV channels, shops, companies, but also me, you and your neighbor. Each of us has access to exactly the same options and, potentially, the same audience.

Alfred Hermida argues that journalism itself has become ambient, and that Twitter functions as an always-on asynchronous awareness system for journalists and the audience alike, creating “new kinds of interactions around the news.”¹⁵ Therefore we might stop thinking of digital journalism as the equivalent of the traditional journalism, where the distinction was clear. “More broadly journalism has been distributed and interlinked more fluidly with citizen communication.”¹⁶

At the beginning of the current year the BBC has asked itself what is the meaning of news nowadays and in the future and has created a detailed and open to the public website¹⁷ where it shares thoughts and ideas on this topic. “This collaboration with the connected audience will transform the BBC’s production

14. Ibid.

15. A. Hermida. 2010. *Twittering the News*. Journalism Practice 4:3, 297-308. Routledge. London, UK.

16. S. D. Reese; L. Rutigliano; K. Hyun; J. Jeong. 2007. *Mapping the blogosphere: Professional and citizen-based media in the global news arena*. Journalism 8:3, 235-261. SAGE Publications.

17. BBC News and James Harding. 2015. *The future of news*. <http://goo.gl/KaASXY>

process. It will be less one team deciding what everyone should watch because now the audience will be our co-producers.”¹⁸

The real question it raises is not who is producing news, but ‘What is information?’ User generated content produces an information overload, a huge amount of content that creates “a mixed of nonsense and news”¹⁹. This could represent an issue for those journalists who want to get meaningful information from UGC. As Steven Waldman highlights, “abundance of voices does not necessarily mean abundance of journalism.”²⁰ In fact, even though “journalists are one of the voices in a noisy information space outside of the formal constraints of traditional journalism” and “in which all the individual pieces of information might seem insignificant, [...] it is the communication as a whole (i.e. the flow) that brings validity and significance to the users.”²¹ The research of meaning in this mixed and booming ambient is afterwards not an issue, but the opportunity to get a sense of the public zeitgeist. And is not impossible: hot topics, trends and most discussed events for example are patterns already easy to identify.

18. S. Kelly. *Click for BBC*. <http://goo.gl/nuj10m>

19. U. Hedman. 2014. *J-Tweeters*. Digital Journalism 1-19. Routledge. London, UK.

20. S. Waldman. 2011. *Information Needs of Communities: The Changing Media Landscape in a Broadband Age*. DIANE Publishing.

21. U. Hedman. 2014. *J-Tweeters*. Digital Journalism 1-19. Routledge. London, UK.

Journalists
and
social media

AN INTRODUCTION TO SOCIAL MEDIA

Earlier in this text I've already touched on the birth and raise of social media, though I still think there is need to think about a proper definition. What are social media? I report here the explanation given by Boyd and Ellison: "web-based services that allow individuals to construct a public or semi-public profile within a bounded system, articulate a list of other users with whom they share a connection, and view and traverse their list of connections and those made by others within the system."²² Apart from the technical definition (web-based) I would like to carefully analyze few aspects, which I find extremely relevant. Social media are services, not product. This means that they solve a need, the one of connectivity. From this comes the definition of social networking, where multiple and interconnected links generate a networked environment, which turns out to be social (it connects people). The last thing I would highlight is the use of the term *traverse*. This word expresses motion, travel across or through, move something back and forth or sideways, as every dictionary would explain. Hence the term emphasizes the dynamic characteristic of these social connections, remarking its evolving behavior.

To better understand how the last decade shaped the digital environment I find important to briefly think over the history of the most used social media. Even though it exists a wide and varied range of different social media platforms, I will limit my choice on Facebook, Twitter, Instagram and YouTube.

Facebook is certainly the most used social networking website and that is why I cannot exclude it from my analysis. It was launched on February 4, 2004 by Mark Zuckerberg with his college roommates. Initially the access was limited to Harvard's students but it gradually expanded first to other colleges, for opening in

22. D. m. boyd; N. B. Ellison. 2007. *Social Network Sites: Definition, History, and Scholarship*. *Journal of Computer-Mediated Communication*. 13:1, 210-230. Blackwell.

2005 to everyone in the world who is older than 13. Registration is required and since it gives users the option to provide several personal information (such as gender, age, living place, educational history, job position, etc.) it has the potential to be the widest census of the online community. However each user has the right to choose whether to leave these information public, available to his circle of friends or private only and we must not forget the presence of many fake profiles, a serious phenomenon not only because it undermines the validity of a census, but because of its dangerous implications.

“As Internet is teaching, conversation rules.”

– Fons Tuinstra, China Speakers Bureau

Then Twitter. The platform is unique for its limitations: users are able to share (and read) messages of 140 characters only. Initially this was due to the vision of a platform built on top of the SMS technology. Even if things have changed (tweeting with SMS is still possible, but not a core anymore), the 140 limit is still there, reinforcing the idea of the tweet, a short and fast cheep. Shortness and frequency of messages are key elements on this space, which is the most used microblogging platform. Every tweet in facts is seen as a small piece of information, which turns out to be a fragmented version of a blog, where posts are replaced with tweets, many in a day. Another key element is the stream idea when in front of the Twitter interface: since it collects the public information (from the profiles the user follows) and in a big amount, the expected behavior is to follow where the stream goes, rather than reading everything it is published. Again, Twitter puts you in contact with anyone, without the necessity of being accepted (even though it is possible to keep tweets available to approved followers only); this often leads to a much wider network than the Facebook one, or at least different: it's less about the ones we know and more about those we are interested in, whether they are people, VIPs or companies. The platform, founded on March 21, 2006, leaves all

of its content publicly available, marking a different line from other social media: sharing with all, not with the chosen ones. This means that the content is available also to those not registered users, making Twitter a valuable tool for researchers and academics in general. An important step in Twitter's history for this research is the November 2009 change of the text that appears at the top of the home page to invite users to tweet: from "What are you doing?" to the more interesting one "What's happening?." This again highlights the news-oriented heart of Twitter. Co-founder Biz Stone has underlines this direction in a Reuters interview in 2010: "From the very beginning this has seemed almost as if it's a news wire coming from everywhere around the world. [...] I think a Twitter News Service would be something that would be very open and shared with many different news organizations around the world."²³

“All the individual pieces of information might seem insignificant, while it is the communication as a whole [...] that brings validity and significance to the users.”

– Ulrika Hedman

I have not mentioned yet how different social media platforms differs when analyzed under the kind of content it is mainly shared on them. Facebook and Twitter are not specifically bounded to one content type: users share textual thoughts as well as photos, videos or links to other sources. Other platforms however focus all and only to pictures like Flickr or Instagram.

In my analysis I decide to not consider Flickr. The reason is very simple: I'm looking for those virtual spaces where people, no matter their professional

23. M. Cowan. 2010. *Twitter co-founder hopes to create news network*. Reuters. <http://goo.gl/mDJcFN>

skills, post content that reflects their perspective on something, event or topic. Flickr, founded in 2004, has a background in professional photography and many users were making of their profile a sort of online public portfolio, where every posted photo is open to comments and to be favorite by others. Instagram instead, works much more as a social network and even though the main options are the same as Flickr (posting, commenting, favoriting), Instagram is less about the quality and more about the moment. While Flickr was born as a service to publish works, Instagram was the software by which shoot and edit a square picture to share. Founded on October 6, 2010, Instagram is first of all a mobile app. The name itself explains the intentions: a combination of Instant-camera (the app is famous for its processing filters which give a vintage look to all the pictures you take) and Telegram (highlighting the sharing component). Instagram make a strong use of hashtags as well, giving the possibility to mark pictures as belonging to similar topics or specific keywords, helping on the other side to discover others' pictures (and users). As other social media platforms each profile is by default set to share pictures publicly, however it is possible to switch this privacy settings differently. The relations network works as the Twitter one: it's not about mutual consensus (as Facebook), but as a following-followed one. The company has been acquired by Facebook, Inc. in 2012.

Since audio-content based social media platforms (such as *SoundCloud*) are not relevant for this work, the only missing format in this analysis is video. There are mainly two video-only based platforms to be widely used right now, namely YouTube and Vimeo. Also in this case I have chosen to pick one only: Vimeo in fact is meant for professionals. Quality of the video uploaded and specifications such as camera and software used are common and somehow expected on this platform. For these reasons, is almost not at all a UGC space, making the content uploaded less reflecting a point of view of the components of the mass. YouTube instead is the main platform used to share and to watch audiovisual material online. Founded on February 14, 2005, it has been bought by Google in 2006. The content shared on

YouTube are of mainly three kinds: user-generated, video-blogs, large production. The latter category could be expanded into the music video industry (YouTube as the new MTV) and the TV one (educational videos, news or other contents). It is interesting for this research the impact that this platform got into the news system: a *Pew Research Center* study²⁴ reported the development of “visual journalism”, in which citizen eyewitnesses and established news organizations share in content creation. As concluded by the study, it is undeniable the role YouTube is playing as an important platform by which people acquire news.

24. Journalism Project Staff. 2012. *PEJ: YouTube & News: A New Kind of Visual Journalism Is Developing, but Ethics of Attribution Have Yet to Emerge*. Pew Research Center. <http://goo.gl/kjsP30>

SOCIAL MEDIA IN THE NEWS SYSTEM

What flows on social media platforms is to all intents and purposes information. In light of this sentence it is interesting to read what in 2009 Richard Gordon pointed out: “an increasing amount of content shared on Facebook and Twitter consists of Web links that search engines cannot see or index. This poses for Google the most serious threat yet to achieving its corporate mission: ‘to organize the world’s

“The news, as lecture, is giving way
to the news as a conversation.”

– Tom Curley

information and make it universally accessible and useful.’”²⁵ Gordon made a very good point, which becomes more and more significant as time passes. However I am lucky enough to report an interesting development that Google gained while I am writing this: on February 5, 2015 Twitter “has struck a deal with Google Inc. to make its 140-character updates more searchable online”²⁶.

This news is not only important for Google as a step towards the accomplishment of its mission, but it is especially relevant in light of the latest *Edelman Trust Barometer* (it is an annual “Trust and credibility survey” that is taken at a global level). The 2015 report²⁷ informs that online search engines, like Google, have globally overtaken traditional media as the most trusted news source. What does this mean for my research? Even though Google doesn’t actually report on news, online search engines also offer the advantage of accessing

25. R. Gordon. 2009. *Social Media: the ground shifts*. in *Let’s Talk: journalism and social media*. Nieman Reports 3:4. The quoted Google mission is accessible here: <http://www.google.com/about/company/>

26. S. Frier. *Twitter Reaches Deal to Show Tweets in Google Search Results*. Bloomberg.com <http://goo.gl/uJh2c6>

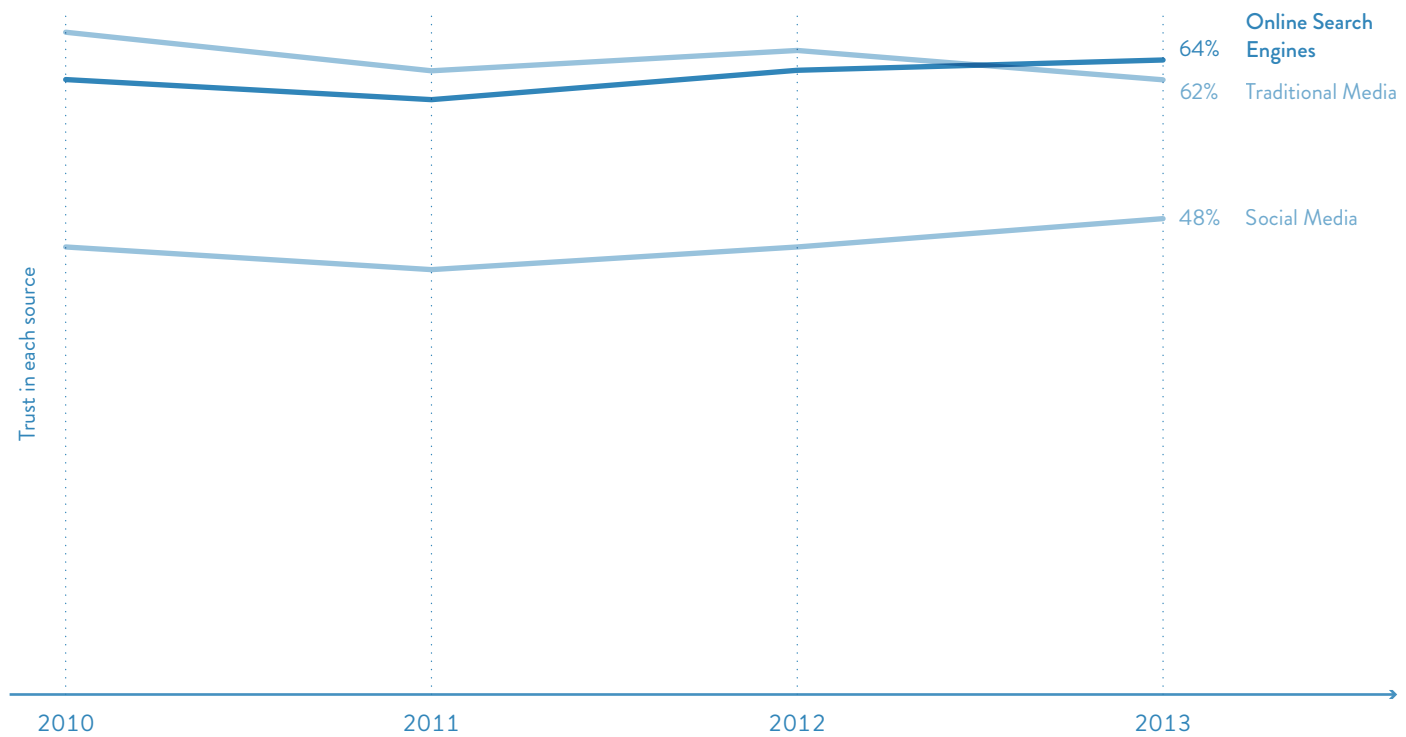
27. *2015 Edelman Trust Barometer executive summary*: <http://goo.gl/6u0UV3>

multiple news sources and viewpoints on a specific story, an advantage that's difficult to match by traditional news networks, newspapers and magazines. Therefore the fact that Google will soon start to show real-time tweets together with its news and websites in its search results page, puts a significant importance to the value of UGC within the information context. This highlights again that social media are a big and pertinent component of the information and news flow of the present of the future.

By giving to users the power to share their own content with a broad audience, part of the role of journalists ceases. Before the raise of blogs and social media the role of a journalist was to fill the gap between information and audience, through a broad communication. This bridge, however, is less and less needed since each user is able to share information potentially at the same level as a news organization. It is thus time to rethink the role of journalism in this new information arena.

After many academics and scholars thought about the role social media should

[8] Trust in each source for general news and information. *Edelman Trust Barometer*. 2015.



have for journalism, news organizations finally started to think of the web as a place rich of new opportunities.

One of the first to be used and of wide adoption, has been the use of social media to increase the popularity of the news organization website. In this case the communication was still top-down driven and the focus was on the presence of the organization on the social. Many researches in fact have shown that users spend much more time on social media websites rather than news organization ones: “users tended to dive into these sites more deeply – often visiting them more often and staying longer. For instance, the average reader spends 20 minutes a month on the New York Times website, compared with seven hours on Facebook.”²⁸ Having a presence on the social media platforms has therefore become an undeniable part of a news organization existence online: posting links to their articles on Facebook or Twitter became a common practice among news organizations. Data confirm the trend:

“Social networks are driving an increasing percentage of the traffic to news sites, beginning to rival search engines like Google as sources of referrals to news stories. Facebook reported that the average media site saw referral traffic from Facebook more than double in 2010. News websites got 9 percent of their traffic from social media such as Facebook and Twitter in 2011, about a 57 percent increase over 2009.”²⁹

Hedman says “there are three dimensions in journalists’ social media use: surveillance and information, organizational demands, and networking and branding.”³⁰ The latter was actually the first taken into account by news organizations: promot-

28. S. Waldman. 2011. *Information Needs of Communities: The Changing Media Landscape in a Broadband Age*. DIANE Publishing.

29. Ibid.

30. U. Hedman. 2014. *J-Tweeters*. Digital Journalism 1-19. Routledge. London, UK.

“Individual users have the power to dictate which information is important and should spread by the form of retweet, which collectively determines the importance of the original tweet. In a way we are witnessing the emergence of collective intelligence.”

– Haewoon Kwak

ing content and distributing news stories appeared suddenly an opportunity for news in the digital era: more views means of course more profit. But as part of the Web 2.0 culture, social networks provide an opportunity for connecting with people. As social media has shown, people have the power to increase a content popularity, therefore the attention news organizations started to give to social media is understandable (also from economic perspectives). On the same note Waldman points out that “Decentralization and universality [...] insured that the Internet and the web would revolutionize not only the dissemination of news and information but how it was gathered and packaged and by whom. This would radically democratize publishing, make ‘sharing’ an essential fuel to the new media.”³¹

Those known as ‘social media buttons’ made their appearance all over the web on any kind of page and topic, news websites of course included them as well. Those buttons allow each user to like, share, tweet, ... the selected page on their social media profile: every other user in the network can then see the content linked in the platform, increasing exponentially the reachable audience. Their presence work as an invite to the reader to share the article, as the article itself was saying: “If you

31. S. Waldman. 2011. *Information Needs of Communities: The Changing Media Landscape in a Broadband Age*. DIANE Publishing.

like what you are reading, it's worth sharing me with your network.” Most of those buttons include also the possibility to check how many appreciations and shares the page got. Even though for the user is not always a good advertisement (it is indeed common to see zero social activity), those kind of data are definitely useful to the news organization which can use them to get a clear idea of which articles are good and which ones are not.

“The breaking of news is no longer solely going to be the domain of news organisations. What has to be, though, is that role of journalism. Because, in a world where everyone can report on news, there is a lot of noise.”

– Alexis Ohanian, Reddit

Social media brought to news websites even more: public comments. “Online comments are as much about people communicating and interacting with each other, as they are just reacting to a reporter’s story.”³² Being able to comment to a news article is in fact the starting point of giving to the audience a place next to the published content. Because allowing a user to publicly comment an article, brings together other things: more power to the users’ opinion, it exposes both the content and the author to public discussion, gives users the possibility to show flaws or missing details in the story and creates an arena for open discussions. In these terms, an article’s life does not end with its publication, it’s rather the birth and from the moment that it’s online it grows up with comments, shares likes and discussions.

It is important to not forget how Facebook took this opportunity by releasing *Face-*

32. P. Grabowicz. 2014. *The Transition To Digital Journalism*. The Knight Digital Media Center, University of California.

book Connect in 2008. The feature, also known by users as *Log in with Facebook*, gave news organizations the option to let readers comment directly with their Facebook profiles rather than creating a new one for every different website (thing that was keeping many users away from joining and willing to set false names). “Everyone is looking for ways to make their Web sites more social”, said Sheryl Sandberg, Facebook’s chief operating officer about Facebook Connect. “They can build their own social capabilities, but what will be more useful for them is building on top of a social system that people are already wedded to.”³³

For their democratic value, social media are a space in which every user, every profile, has the same options and opportunities. That means that journalists are not anymore on top of their audience, but they now share the same living space. A virtual space that is more and more parallel to the reality, where “being online is rapidly becoming equivalent to being part of society, and the concept of Internet access being as fundamental as access to electricity or telephone service seems less revolutionary every day.”³⁴ “As a result, social sources have become indispensable for the modern professional journalist. Most news organisations expect journalists to be fluent in social media in order to discover and distribute news, and many expect them to hold conversations with audiences as part of the ongoing production process.”³⁵

Connecting with audience started to be of relevant importance for news organizations: it gives the opportunity to build a stronger relation with readers. It also represents what Hedman refers to as organizational demands. Building

33. B. Stone. November 30, 2008. *Facebook Aims to Extend Its Reach Across the Web*. The New York Times, <http://goo.gl/vlfhvs>

34. P. Miel; R. Faris. 2008. *News and information as digital media come of age*. Berkman Center for Internet & Society at Harvard University. Cambridge, USA.

35. S. Schifferes; N. Newman; N. Thurman; D. Corney; A. Göker; C. Martin. 2014. *Identifying and Verifying News through Social Media*. *Digital Journalism* 2:3, 406-418. Routledge. London, UK.

a dialogue with the audience can lead news organizations to know which content and topics readers appreciate most, guiding the choice for future stories and articles positioning.

It is clear how news organizations and journalists are changing their opinion about the value social media bring to the information cycle. Before the Internet the news cycle was pretty much linear: event > journalist investigation > report writing > publication. And even in the most complicated events, the structure could not change since reaching the audience was always a step after the journalist's knowledge. With no press, an event could not be news. But with the Web 2.0 a group of users can start to spread information around the web, reaching a broad audience and reaching in this way unaware professionals. So the information cycle is not defined anymore in its flow: the publication from journalists can be something that comes after the publication of the information from the audience. However, as Miel and Faris highlight:

“Many groups have gained a voice thanks to participatory media, but many others do not participate online or participate in online communities that remain walled off, with their concerns not reflected in the general media space. The potential to use new technologies to improve on traditional inequalities is not being fully realized, which has consequences not only for the media consumption of marginalized communities, but also for their representation in the broader public sphere. As journalists increasingly rely on online sources, populations or ideas that are absent from the online space may as well be invisible.”³⁶

It is very important to not forget this: the Internet created a much more open and free space for our communities, but has somehow increased the gap with those communities who are not online. This is something that the Internet in his struc-

36. P. Miel; R. Faris. 2008. *News and information as digital media come of age*. Berkman Center for Internet & Society at Harvard University. Cambridge, USA.

ture cannot solve, it is governments' job to help those outsiders to join, especially those communities that cannot access the Internet because of the digital divide or due to political issues (see e.g. China and the government censorships of Google and Facebook).

“When everyone, everything and everywhere is connected the possibility for the BBC will be endless.”

– Spencer Kelly, BBC

The most difficult practice to adapt from traditional media to the new digital environment, has been (and still is) regarding the norms of journalism. The shift to a more fast and ‘in-real-time’ journalism heavily shrank the time newsrooms had to gather and verify sources. Ideally there was time ‘till the next edition’ whether it was TV news or the next day paper. Online the rush to be the first to publish a news is even more hard, since as soon as something is shared the Internet society will start to discuss about it and to look for more information making easier for ‘who is late’ to stay away from the public discussion. Having such a short time available often puts speed before quality, impacting one of the most important norms in journalism: telling the truth. As Farida Vis states: “verification of information is one of the cornerstones of journalism” and it is something that cannot be forgotten.

NEWS SOURCES

Imagine an explosion in a building. Chaos starts. People around are shocked. Now imagine that each of those witnesses that are staring at flames and smoke have a camera in their hands, a video recording camera. And in the same device they have instant access to a platform that let them share with friends or even with strangers what they are witnessing. Imagine the value that all this content has for journalists and newsrooms in a breaking news reporting. Actually, there is no need to imagine it, because this is what the combination of a smartphone, Internet connection and a social media platform does. Every person, provided with these three things, can be a relevant news source.

“A few months ago, I was given an additional duty in my job description: ‘Spend meaningful time on Twitter and Facebook’. That’s something I never thought I’d see.”

– Courtney Lowery

There is no doubt then that social media platforms and UGC are highly important for journalists. As a matter of fact there are many cases in which Twitter has been used as a source for breaking news reporting and as an important step in many journalists’ workflow during their researches.

Having this content accessible and ready to use has many other characteristics that I find necessary to highlight. UGC published through platforms such as Twitter, Facebook or Instagram, has often a geolocation indication embedded. Users can of course disable this feature for privacy concerns, but when they find their location to be an important aspect of their content, they are willing to share it, adding trustfulness to the published information. Such kind of content, whether is a



[9] Egyptians use their mobile phone to record celebrations in Cairo's Tahrir Square during the Arab Spring, 2011. Mohammed Abed/AFP/GettyImages.

text information, a picture or a video, contains a specific point of view on what is happening: it can be a way to look at an event or the perspective from which the footage was shot. Having access to multiple different points of view (and the sum can reach thousands and millions) is the only way to get a bigger picture, less influenced by one person's own biases. Again, another addition is the speed and zero-cost at which this information can be gathered: posts happen in real time and

“Social media raises a number of issues about how eyewitness material should be used and credited and, perhaps most importantly, what kind of checks should be made to ensure veracity.”

– Steve Schifferes

are suddenly available to be consumed, for free, by everyone on the same platform. The journalist's research for witnesses is much easier and faster and can be done at distance, without the costs and the long time required to find relevant (offline) sources. Literature already proves the relevance of such kind of approaches, in particular “Distant Witness” and “Watching From Afar” titles say a lot about it, where they explain how breaking news reporters made use of social media content to follow the evolving of the 2010 Arab Spring protests.

As Bruns and Burgess point out, social media, Twitter first, are technologies able to combine convergence of social networking, content production and information sharing. For them, “Twitter is both a social networking site and an ambient information stream.”³⁷ Many are the examples of the relevance of Twitter as a news source, but the simple decision by Obama's team to mark his re-election on Twitter

37. A. Bruns; J. Burgess. 2012. *Researching News Discussion on Twitter*. Journalism Studies 13:5-6, 801-814. Routledge. London, UK.

is emblematic of how the service has become part of the media landscape. Social networking sites or microblogging platforms' content becomes extremely relevant for breaking news reporters especially in two specific cases: when it's impossible for journalists to reach the location of interest (because it's dangerous or forbidden) or when the location is unknown. The UK riots news story is a good example of the latter.

Of high value and using the same approach is the broadly reported work done by Andy Carvin, "the man who tweets revolutions"³⁸ as *The Guardian* newspaper referred to him, a "living, breathing real-time verification system"³⁹ as dubbed by *The Columbia Journalism Review*. *GigaOM*'s Mathew Ingram says that Carvin's "ability to use Twitter as a real-time, crowdsourced 'public newsroom' was a model for what more journalists should be doing during breaking news events, and a great example of what Guardian editor Alan Rusbridger calls 'open journalism'."⁴⁰

Unfortunately, UGC is not a secure source of information. Firstly because social media networks are a noisy space, full of worthless and meaningless content (at least not relevant as a source, especially when the content is personal and out-of-topic). The capacity to filter is therefore a priority in journalists' workflow who investigate a story through social media. There is need to organize the enormous quantity of shared content in order to keep only the relevant one. A second reason is that in an ambient where there is no authority, every user is able to share his own thoughts whether they are true or not. Within trending topics the number of fake news (such as photoshopped images) and the spread of wrong rumors raise exponentially. Yet the verification of UGC becomes critical for breaking news reporters.

38. J. Kiss. September 4, 2011. *Andy Carvin: the man who tweets revolutions. NPR's media strategist has become a leading player in the breaking news business.* The Guardian. <http://goo.gl/fzISHP>

39. C. Silverman. 2011. *Is This the World's Best Twitter Account? Meet Andy Carvin, verification machine.* Columbia Journalism Review. <http://goo.gl/UvV2JH>

40. M. Ingram. 2014. *Andy Carvin, a pioneer in using Twitter for real-time journalism, joins Omidyar's First Look Media.* Gigaom. <http://goo.gl/OylPp8>

The spread of rumors is indeed a serious problem for journalists. Those (not yet verified) stories born on the Internet and they start to disseminate very fast among users across the world. Journalists need to deal with this fast and out of control flow, which ask for their skills to be verified. In a *TEDTalentSearch* 2013 talk⁴¹, Farida Vis explains her work on the analysis of rumors concerning the UK riots case on Twitter. As she introduces, it's about "the role rumors on social media play in a crisis situation" describing that "accurate, up-to-date information is absolutely crucial." In her talk she presents the work done together with Rob Procter, Alex Voss and the *Guardian Interactive Team* "How riot rumours spread on Twitter"⁴². The focus was to understand better the life cycle of rumors. After manually coding the tweets and mapping them on a timeline, they showed that within an hour from the spread of the rumor there actually is opposition to the rumor, starting from one single tweet that points out how it cannot possibly be happening. Vis concludes showing that "social media is actually really effective itself at debunking these rumors."

Newsrooms acknowledged they are facing these difficulties when dealing with social media as news sources. As a matter of fact not only scholars and academics took in consideration the issue, but social media platforms themselves and news organizations started to create guidelines to help journalists in their use of such tools. Twitter for example published a whole *News*⁴³ section on its *Media* website (a space on which it features practices and stories for a professional use of its social platform). Among the published articles the most interesting ones at the day I accessed the website (February 22nd, 2015) are "Techniques for covering breaking news events" and "News: The impact of tweeting with photos, videos, hashtags and

41. TEDTalentSearch. 2013. *Farida Vis: Social media and the life cycle of rumors*. <http://goo.gl/tsiYrs>

42. The Guardian. 2011. Reading the Riots. *How riot rumours spread on Twitter. Analysis of 2.6 million tweets shows Twitter is adept at correcting misinformation - particularly if the claim is that a tiger is on the loose in Primrose Hill*. <http://goo.gl/rgwsYf>

43. <https://media.twitter.com/news>

links.” Those are just two examples of how self-aware Twitter is when it comes to its use in the journalistic field. Even though compared to Twitter it is still less used from journalists, Facebook as well has developed some sort of guidelines for those professionals: on its developers section there is an entire page about called “Media: Unleash the power of stories on Facebook”⁴⁴ that features links such as “Best practices for journalists on Facebook” or “Update breaking news stories with new content as they develop.” It has even verified (with the blue badge that certifies the authenticity of a profile or page) the “Journalists on Facebook” page, another sign of the significance of the use of the platform for journalists. On the same note the social media has dedicated an entire website called *Facebook Media*⁴⁵ which presents itself with the tagline “Explore how public figures and media organizations are using Facebook in extraordinary ways.”

From the news organization side, instead, the BBC has often been ahead to many competitors and its “News: Social media guidance”⁴⁶ it is yet another proof of this. The 2011 update to the guidelines are in fact focusing on the use of social media for its professionals, marking the difference between private life and public figure as representative of the company.

In a 2013 talk⁴⁷, Andy Carvin asks his audience: “Why use social media for reporting?.” He then shows some of the most common answers (that he labels as wrong) among which the improvement of Facebook presence and the increase of traffic to the website. “Correct answer: because it can help improve my reporting.” In saying so, Carvin highlights the importance of the use of social media as news sources. His method is based on searches from a eyewitness point-of-view. Interviewed by Mathew Ingram in 2014 about his choice to join *First Look Media*, Carvin

44. <https://developers.facebook.com/docs/media>

45. <http://media.fb.com/>

46. *BBC News: Social Media Guidance*. June 2011. <http://goo.gl/UQutKH>

47. A. Carvin. 2013. *ICFJ Mid East boot camp in Rabat*. Slide presentation: <http://goo.gl/1x4l8x>

says: “Even with some digital news startups over the last several years, social media is usually an add-on to promote their content, to try to get it to go viral, etc. This (joining *First Look Media*, Ed.) feels like a great opportunity to see what it would look like for a news organization to be engaged in collaborating with the public on a more fundamental level across its operations.”⁴⁸ In a later interview Carvin adds: “Not many people are thinking of these places as living and breathing spaces where they can discuss the news directly with the people who are there.”⁴⁹ With this, Carvin explains even better his new vision on the use of social media: not only platforms where to extrapolate useful content, but places of communication, where the reporter actually get in contact with those he might think are relevant sources. User-generated content is the only way we have to measure what the society shares and therefore what the mass thinks. It is the only way to map where and when something is being discussed on such a broad, and at the same time granular, level. This is realistic due to the Internet and to the openness of social media: living in a historical period where a single opinion becomes public information has the potential to see differences and similarities in knowledge and thinking. I think this is an important and relevant change in journalists’ process of source gathering. However it is important to not forget the limits of this approach. Having access to such broad information, together with the risk of getting into fake statements or misunderstandings, could lead to the idea that everything that is online is everything that exists. That is definitely false and it is important to remember that out there, on the ground, there are many events and many voices that do not find their place in the digital environment. Acknowledging this and keeping it in mind, will help researchers to not draw conclusions too soon, but it will rather give them new and unthought directions to follow, without the risk of don’t take into account the ‘forgotten communities’.

48. M. Ingram. 2014. *Andy Carvin, a pioneer in using Twitter for real-time journalism, joins Omidyar’s First Look Media*. Gigaom. <http://goo.gl/OyIPp8>

49. M. Ingram. 2014. *Andy Carvin launches social-media reporting team for First Look*. Gigaom. <http://goo.gl/OnXBhf>

Some have even speculated that social networks will supplant news websites as the place where people get news. This is not really the case, at least not in the upcoming years. In fact, technology and “the automation of many tasks in news

“In a world where everyone can report on news there is a lot of noise, and the journalist’s role is now more important than ever to find the signal in all of that noise and help tell a story with authority and with context that helps all of us understand what the hell is actually going on.”

– Alexis Ohanian, Reddit

content analysis will not replace the human judgment needed for fine-grained, qualitative forms of analysis, but it allows researchers to focus their attention on a scale far beyond the sample sizes of traditional forms of content analysis.”⁵⁰ It is clear the necessity of tools to help the analysis in an information overloaded environment, but on the other side is much more important the work of someone who can make order throughout the noise. This is still an open opportunity for journalists to do at best their jobs in this new ambient. Tim Berners-Lee, of the *World Wide Web Consortium*, shares this thought when he says: “The problem of how to distinguish good information from bad, that problem has been with us since we started communicating... So even though we have a new technology where information comes to us instantly over the wires... the art and science of journalism is really important.” Also the co-founder of *Reddit* Alexis Ohanian agrees that “In a

50. I. Flaounas; O. Ali; T. Lansdall-Welfare; T. De Bie; N. Mosdell; J. Lewis; N. Cristianini. 2013. *Research Methods in the Age of Digital Journalism. Massive-scale automated analysis of news-content- topics, style and gender*. Digital Journalism 1:1, 102-116. Routledge. London, UK.

world where everyone can report on news there is a lot of noise, and the journalist's role is now more important than ever to find the signal in all of that noise and help tell a story with authority and with context that helps all of us understand what the hell is actually going on.”

The needs

THE TARGET: ANALYTIC JOURNALISTS, BREAKING NEWS REPORTERS, AND ACADEMICS

While researching on the relation between journalists and social media, I ran through different ideas and different professions. It appeared clear that assuming journalists would use a tool to explore social media is not totally true. In fact even though many journalists believe in the value of UGC, not as many would actually use it in their workflow to find new perspectives or relevant news sources. For this reason I understood it was necessary to identify a more specific target. The journalist profession is indeed very broad and different professional roles cover a wide range of different genres in journalism.

The main outcome of the research done is that the journalism category, which makes more use of social media, is the one dealing with breaking news. A breaking news reporter is clearly continuously focused on being aware of things in the moment they happen. Especially in the era of digital journalism there is potentially no delay time between event and publication of the news, there is no more possibility to wait and investigate till the end of the day when the paper goes to print. Social media, with the attitude of posting real-time things, are the place where something is being told (or shown) publicly as soon as it happens. This is of significant value for breaking news, especially for the unplanned happenings (in those cases where the press is not present at the event).

This genre of journalism is also known as spot news and it is about reporting of events as they occur. “Breaking news refers to events that are currently developing, or ‘breaking’. Breaking news usually refers to events that are unexpected, such as a plane crash or building fire. Breaking news can also refer to news that occurs late in the day, close to a news outlet’s usual deadline.”⁵¹ It is clear, then, how important is for this journalists to be informed on something as fast as possible. Having the

51. T. Rogers. *Journalism, the basic of reporting*. Definition of *Breaking News*. About.com

possibility to browse social media in an easier way could help them to spot relevant information of something that just happened, in order to enrich the content of the breaking news article.

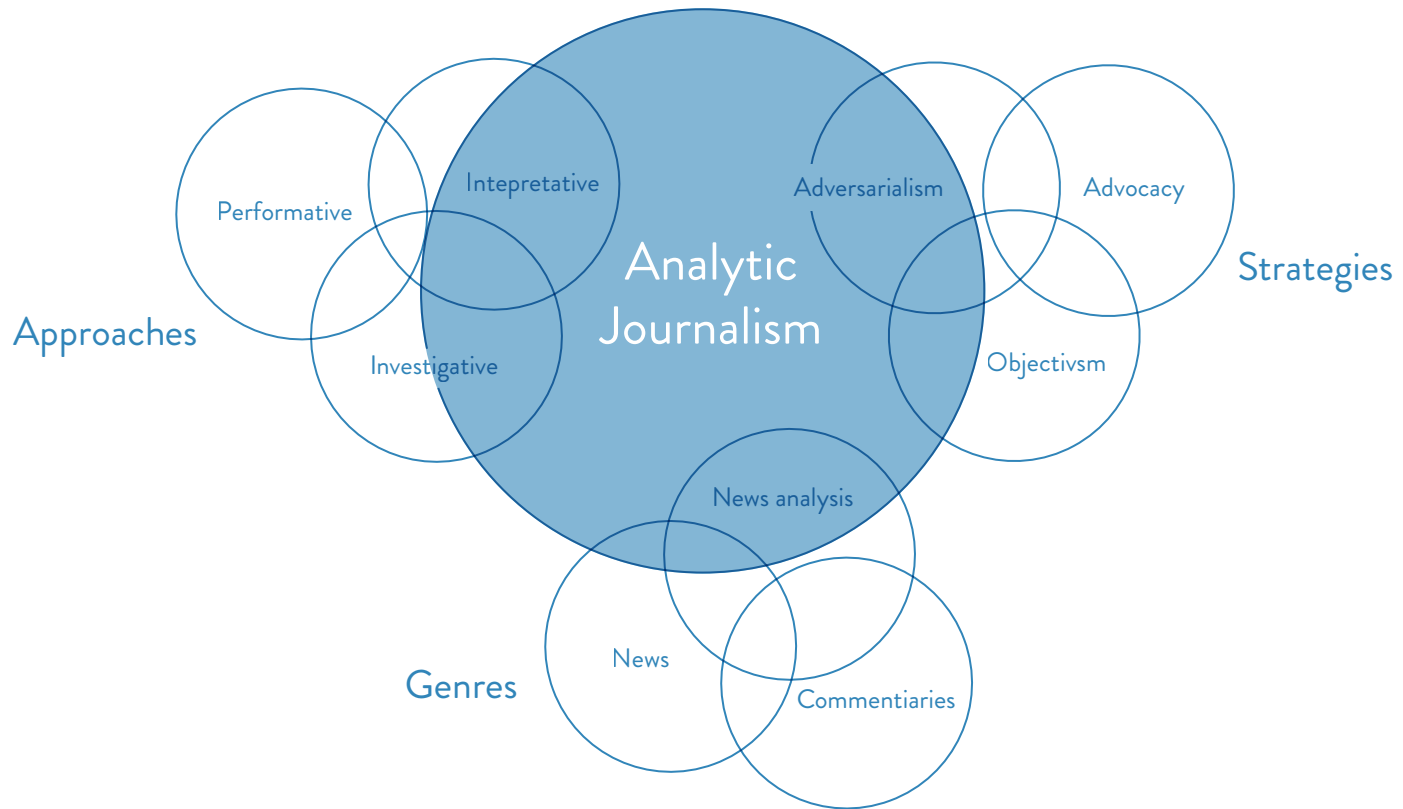
Another genre is the analytic journalist. This role works on in-depth coverage of a specific story and for this reason it researches broadly and deeply on it. Professor Svith Flemming in a university lecture explains the position of analytic journalism in context with other genres, approaches and strategies. The analytic journalist, he shows in a diagram, works with an interpretative and investigative approach and his position is very close to the adversarialism and objectivism. Even though I personally do not believe it exists any kind of objectivity (especially when it comes to tell a story), it is interesting to see the interpretative approach, which is the most important when dealing with big amount of information, where interpretation lead to one way and not to another.

I find the Wikipedia definition of this genre quite explicative:

“Analytic journalism seeks to make sense of a complex reality in order to create public understanding. It combines some aspects of investigative journalism and explanatory reporting. Analytic journalism can be seen as a response to professionalized communication from powerful agents, information overload and growing complexity in a globalised world. It aims at creating evidence-based interpretations of reality, often confronting the dominant ways of understanding a specific phenomenon.”⁵²

Comparing analytic journalism with news reporting, De Burgh says: “news reporting is descriptive and news reporters are admired when they describe in a manner that is accurate, explanatory, vivid or moving, regardless of medium. Analytic journalism, on the other hand, seeks to take the data available and re-

52. Wikipedia. *Analytic Journalism*. Accessed on February 27, 2015.



[10] Analytic Journalism within the Dimensions in Journalism. S. Flemming. Aarhus Danish School of Media and Journalism. Unpublished lecture. (redrawn)

configure it, helping us to ask questions about the situation or statement or see it in a different way.”⁵³ He continues his analysis saying: “the duties of today’s journalist can be divided roughly into three basic functions: Hunter-gatherer of information; Filter; and Explainer. [...] Only in our role as ‘explainer’, as storyteller, do journalists appear to have a reasonable secure position. To ‘explain’, they have to do more than ‘report’.”⁵⁴

The last target group I identify as significant, for the use it may do of social media in its workflow, is the one of academic researchers. Scholars and academics are more and more looking at social media as places of public discussion, where

53. H. De Burgh. 2008. *Investigative Journalism*. Routledge.

54. S. Ross. *Teaching Computer-assisted Reporting on South India* in Nalini Rajan (Edited by). 2005. *Practising Journalism: Values, Constrains, Implications*. Sage. New Delhi, India.

communication among different social group happens. Investigating how a topic spreads online is already the subject of many research studies, either when the goal is to better understand a topic or when instead is more about social behaviors.

As mentioned in the previous chapter, social media researcher Farida Vis is a clear example of how social media analysis became a strategy to conduct social studies. An increasing number of university courses raised on this topic since the analysis of data coming from platform like Twitter gives academics something previously unavailable. Such data are in fact regarding users from all over the world and of a wide range of age and professions, with different cultural and educational backgrounds.

RESEARCH METHODS: FROM EXPLORATION TO VERIFICATION

Even though the identified target groups have different approaches in the way they do their researches online, I will hereby describe some of the used methodologies I have encountered during few months of research on the topic of journalist's needs and opportunities in the digital era. I will not make a distinction in which technique is being used by which group because I believe that even within the same group there is not just one approach and in doing so I would give a wrong-headed vision on the workflow of these professionals.

We can assume, however, that the research process follows a similar workflow among the professional, whether they are academics or journalists. As explained by the diagram, any research starts with the topic choice. A topic coincides with the subject of the research and can be chosen, for example after a preliminary (and sometimes unconscious) observation. A subject can also be given, imagine in a newsroom that a specific news event is assigned to a journalist for a news story.

The first phase consists of *exploration*: the researcher looks around, observes and tries to absorb a big picture of the case. In this phase and in the next one, the researcher formulates hypothesis, which need to be verified. After it comes the writing phase, which does not need to be strictly in a textual form (journalism relies now on many format, for example the interactive visualizations used by data journalists). Finally the researcher-author publishes of the results in his news article or scientific paper. Especially in the case of journalism, this process does not end with the publication: monitoring the engagement of the story and getting in touch with the audience is becoming an important part of journalists job.

To get more into the details of this process, the exploration phase relies on hunting information. It is the starting point of a process in which the digital researcher explores the online environment while looking for answers. Moreover it is important that the researcher gets a good understanding of the topic he is working on; however, when the amount of information is too big to be possibly handled,



[11] Journalists' workflow.

there is need to filter. Filtering is an essential step in this phase and it leads to the pieces of information that needs to be further investigated.

The second part of the research process is *investigation*. Even the exploration method can look quite alike the one of the investigator, who looks around in search for clues (which could be scoops for the spot reporter or answers to a research question in the academic's case).

Talking about social media, researching on a specific topic is made easier by the use of hashtags. A hashtag is a hyperlink which can be created by any user in any part of his/her post just by using the *octothorpe sign* (ASCII code U+0023) commonly known as '#'. Placed at the beginning of any word, it transforms it into a link, which brings to a page that gathers together all public posts using that keyword. The introduction of this technology, nowadays supported and largely used by most of the social network platforms, is an interesting method in the research on topics that use specific hashtags. It creates a sort of code, where a specific hashtag identifies a post as pertinent to a certain category. From a social perspective it is interesting to see how and when does a hashtag gets created, by who and why a specific one starts to being used more than another. Quite often one topic is supported by more than one hashtags (it might be relevant to see in what they differ from each other and if there are any changes across countries when the topic is related to a global discussion). Some platforms, such as Instagram, started to abuse the hashtag, collapsing entire sentences in one-word hashtags (spaces and other punctuation signs break a hashtag link). This attitude is generally nonsense for the usefulness of a link, so even though the hashtag exists and it works, it's like it has never been created (if it is used only once, the link won't show more results). In some cases, of large use again on Instagram, the hashtag is used to literally mark content with widely searched terms, helping users to discover and being discovered.

It is important to note that the hashtag technology it is used only on social media platforms and not by news articles. That means that researching on different frameworks need to take in consideration these differences. I would say that basing

a research only upon the use of a hashtag might be distorted. In fact it has been shown by many studies that associated to a single event there might be different hashtags in use. Not to mention all those posted contents that do not use the searched hashtag or no hashtags at all: those will never be part of the results in the search. However this could be a choice based upon a different perspective. In fact a hashtag could be more specific than a word to identify an event. Therefore looking for the hashtag would generate a smaller, but more specific result.

If the hashtag or keyword in general could be the starting point of a research, gathering the information needed, the relation between who generated that content and the audience might be relevant in the most cases.

For this kind of associations both Twitter and Facebook make use of the *at sign* (ASCII code U+0040) known as '@'. However a Facebook user uses it differently from a Twitter one. The difference is also in the terms use: on Facebook we *tag*, on Twitter we *mention*. The first one writes it before typing a friend's name, when the system recognize enough characters to match a friend, the user need to selected one of the suggested matches. After this process the name becomes a link (which brings to the friend's profile) and the @ disappears. On Twitter instead, every user is defined by a certain username which contains by default the *at sign*. That is that when mentioning a user the @ will be part of the post even when it's published. If the specific name of the user does not match, the combination '@+word' will not generate any link, meaning the mentioning was not successful (the interface of Twitter suggests matches as well to help the user to not make it wrong).

So, from a social point of view, relations between users are important and can be measured in different ways: is users mention each other or, even first, if it exist any connection between them in the network.

On Facebook it's about mutual relations: two users are connected if they both agreed on it (the relation is then called *Friendship*). On Twitter instead it's about *following*, generating three possibilities for each profile in relation to another: A follows B, A is followed by B, A follows and is being followed by B.

For its openness, Twitter is much more used as a research tool both by journalists

and academics. In fact for all users on Twitter is possible to know the number of followers and follows, which gives an easy way to measure the importance of a user on the platform: is this an important node in the network? Then it is highly probable that it will score a high number of followers. Is this an active user? It's just a matter of seconds to check when his last post was published or the number of tweets posted since his membership.

Another technique used when trying to get more information on a specific topic, especially within a specific location, is the use of groups of users. Gathering together few trusted sources, chosen for their expertise in a certain field or for their geographic position, helps the researcher to collect in one place all the information that might be more significant. It serves as a sort of bubble in the noisy and cluttered space that UGC creates. In this case the researcher-explorer is looking for updates on an event from those who knows more than him, for this reason he needs to separate these voices from the others.

In the previously cited work done by Andy Carvin on the Arab Spring this method was widely employed both on Twitter and on Facebook. On Twitter every user can add those who is following to specific lists. Andy Carvin did it by location, so by switching to view the tweets from one list or the other, he was getting insights on the situation in the different countries involved in the protests. Filtering social media content by location is what Carvin uses to cover regional issues. Location can be the place a user says to live in (public on Twitter, of user's choice on Facebook) or the one of the specific post, indicating where the user says to be in the exact moment of the post (public in both Twitter and Facebook if the post is set to public). This means that the content in analysis is strictly connected to a specific place, which can be of the order of a city, a region or a nation.

Equivalent to the advanced search of Twitter, Facebook offers *Graph Search*. Introduced on the social platform in March 2013, the semantic search engine is designed to give answers to user natural language queries. That means that in the search field we can normally write the sentence that ask what we are looking for, such as "Friends of my friends who live in Cairo, Egypt who are activists

and live in Cairo, Egypt” as Andy Carvin shows⁵⁵ as an example of his method. Advanced searches, whatever the platform is being explored, follow the principle of the filter. Filtering is an effective way to make some clarity in the noise: after all being able to select part of a big quantity of information based on similarities is what we do every day when trying to make order. We establish mental categorization that let us group things. In the case of UGC there are usually different options to filter results: words used (individually or together), words not used, language, user’s location. Recently Twitter has open its search results to the whole Twitter history, adding time to its search filters, letting you define a starting and ending date to focus the search. Another approach used by Carvin is to look for content “from an eyewitness point-of-view.” That is that the search is driven by words or expressions typical of someone posting a content such as: “WTF” “What the hell was that” etc. Within the filter as an investigatory approach, when it comes to real-time reporting it is more proper to talk about monitor. For real-time reporters keeping up with latest news is vital, since their work consists in publish it as soon as it happens. In

[12] Mark Zuckerberg launching Facebook Graph Search. 2011.



55. Andy Carvin. 2013. ICFJ Mid East boot camp in Rabat. Slide presentation: <http://goo.gl/1x418x>

doing this one of the most used tool is *TweetDeck*. Employed not only by journalists, but also by individuals and companies, it makes the monitoring component of the workflow very easy. The tool is a Twitter product that helps in monitoring, managing searches and tweets scheduling in one interface. I would call it the pro-version of twitter. It presents an interface divided in multiple columns, each customizable with the function we want it to fulfill (there is no limit to the number of columns). Columns usually consists of *timelines*, the name Twitter gives to streams of content filtered by specific lists or hashtags or mentions. Many news reporters use TweetDeck in two ways. One is to get informed (there is also an option to turn on alerts) when something new happens. For example they create alerts based on specific words used, such as “Earthquake”, to ensure they are aware if something ‘breaking’ happens. The second use is the one that enables further monitor on a specific story they are following. The principle is the same, but the use of custom lists become more prominent and the key terms followed might be multiple. The columns interface gives the opportunity to keep everything in sight in real time, since the software updates every column constantly with no need to refresh or to load more. Again, the Internet is a pool of tools that helps in filtering and gathering together UGC from different social media platforms. Browsing images that were uploaded from a specific location is easy with *Geofeedia* or *Ban.jo*. “Banjo is the world’s largest collection of social signals organized by time, location and content. We index and curate breaking news and events faster than any organization on the planet so that you have an all-access pass to anything, anywhere”⁵⁶ says the about section of the platform. Geofeedia on the other hand presents itself with the pay-off: “Search, archive & analyze real-time social media content across multiple sources, from any location in the world, with a single click.”⁵⁷ While Banjo look more as a curated content with professional footage, Geofeedia (which offers subscriptions) provides a view of UGC images from Twitter, Instagram, Facebook, Flickr, YouTube, Picasa

56. <http://ban.jo/about/>

57. <http://geofeedia.com/>

and others. The service offers also a dashboard view and other useful tools such as the possibility to draw an area on a map from which you want to view the shared content. The website has also a specific page of presentation for the media: “Breaking news doesn’t wait for hashtags. Access real-time photos and videos from any location in the world, across multiple social media sources, with a single click.”⁵⁸

For both journalists and researchers, the *verification* of UGC represents an important step in their workflow. For breaking news reporters especially, who deal with small amount of time, verification needs to happen efficiently and fast. Josh Stearns, who won the 2011 *Storify of the Year* with his (real-time via social media) report ‘Tracking Journalist Arrests at Occupy Protests Around the Country’ says: “As the tools we use to report online continue to shift, we need verification to keep up.”⁵⁹

I find significant for this chapter to report part of the text on the back cover of the “Verification Handbook” edited by Craig Silverman and released by the *European Journalism Centre* in 2014: “Social networks provide us with a pool of story leads, witness photos and videos. But how do we make sure that they are trustworthy, authentic, accurate source for news coverage?” It continues in the introductory pages: “In a crisis situation, social networks are overloaded with situational updates, calls for relief, reports of new developments, and rescue information. Reporting the right information is often critical in shaping responses from the public and relief workers; it can literally be a matter of life or death.”⁶⁰ Claire Wardle, research fellow at the Tow Center at Columbia University, explains her view on verification as a process: “Unfortunately, people often see verification as a simple yes/no action: Something has been verified or not. In practice [...] verification is a process. It is rel-

58. Geofeedia website. *Industries* page, *Media and Journalism* section. <http://goo.gl/kl8RqN>

59. Josh Stearns. 2013. *Verifying Social Media Content: The Best Links, Case Studies and Discussion*. <http://goo.gl/3kb7JX>

60. C. Silverman (Edited by). 2014. *Verification Handbook, A definitive guide to verifying digital content for emergency coverage*. European Journalism Centre. Maastricht, The Netherlands.

atively rare that all of these checks provide clear answers. It is therefore an editorial decision about whether to use a piece of content that originates from a witness.”⁶¹ Overall the verification process roughly consists in this checklist: provenance, source, date and location. When all of these have a certain answer, it is luckily that the UGC is trustful. At least, this can be the new starting point for further investigations (such as directly contacting the source) in pursuit of the truth.

“Applied to the discipline of verification, a new literacies approach suggests journalists adopt a more collaborative method to determining the truth that, in theory, could be reached through an iterative process played out on networks such as Twitter.”

– Alfred Hermida

There are different methodologies used by professionals to understand if a picture is trustful or not. Every picture taken with a camera or scanner incorporates by default some additional information in the form of metadata tags, called EXIF (Exchangeable image file format). These contain information about brand and model of the device used, date and time of the shot and in some cases even location (among others). Accessing these data is of course very useful for the verification process since it can answer many questions: When was the picture taken?, At what time?, From which device was it taken? Then guiding the researcher in checking if the material matches with the known information. Extrapolate the EXIF data is relative-

61. Claire Wardle; Craig Silverman (Edited by). 2014. *Verification Handbook, A definitive guide to verifying digital content for emergency coverage*. Chapter 3: Verifying User-Generated Content. European Journalism Centre. Maastricht, The Netherlands.

ly easy: software such as Photoshop or even some websites are able to show those within few clicks. However social media platforms destroy this idyllic method: for privacy concerns Facebook, Twitter and Instagram erase the EXIF information from any content before it is published.

For this reason, professionals have to make use of other tools to workaroud this problem. The method consists in looking for similar pictures. In the most lucky case this will show that the same picture was previously published somewhere outside the social media, giving therefore higher chances to reach the EXIF data related. For doing this two free tools are being used: *Google Search by Image* and *TinEye*. Both of them let a search begins with an image, rather than a textual query (Google call this 'Reverse image search'). There are two opportunities: look for 'Other sizes' or for 'Similar images'. Other sizes will look for the exact same image (similarity value set to very high) in different sizes, giving the possibility to encounter higher resolutions. Since an image cannot be increased in size without loosing in definition, usually the largest size match is actually the original source or, at least, the closer to the original that the search engine was able to find. That means that opening the web page containing the found image will lead to the place on the web that first uploaded it. In this case many are the benefits: the posted picture is not original, the found match could contain EXIF data useful for further researches, looking for the date of the post can be enough to validate it. The other option is a bit less useful in the verification process, but it could still help in some cases. Checking similar footage could lead to browse through images of the same event, useful to make comparisons and understand if something was clearly manipulated.

Another technology useful for image verification is the much less known *Image Error Level Analyser* developed by Neal Kravetz. The algorithm (ELA) "identifies areas within an image that are at different compression levels. With JPEG images, the entire picture should be at roughly the same level. If a section of the image is at a significantly different error level, then it likely indicates a digital modification."⁶²

62. Hacker Factor. 2012-2015. *Tutorial: Error Level Analysis*. <http://goo.gl/Gd8Bjs>

Even though the description and some results show the effectiveness of using this tool to discover image manipulations, it is in my opinion a difficult method and it is likely to require the help of an expert in such analysis. For its unhandiness I therefore do not think this is a relevant tool for journalists, on the other hand researchers, who work in a less timed process, could reasonably employ it.

Sometimes an image may be authentic, but it could be inaccurately labeled. That is the case, for example, when an older picture is being reposted as something happening in real time. It is therefore important to “confirm that the image is what it purports to be. An authentic image can still be placed in a false context”, says Trushar Barot, assistant editor at the Social Media and User Generated Content hub at BBC News.

When dealing with shared videos, till now there is no tool that works as the reverse image search. A common practice consists in using few frames of the video (among which those used as a thumbnail preview) as queries for an image search. This way results will point to webpages containing the video or talking about the video with some extracted frames. Fortunately enough, being able to fake videos require higher skills than that of images manipulation. Still, videos and pictures together present new issues, as Claire Wardle suggests: “People downloading content from YouTube and uploading it to their own accounts, claiming it as their own, cause other problems. This isn’t a hoax - it’s what is known as a ‘scrape’ - but it means we have to work harder to find the original uploader of the content.”⁶³ Being able to find the first uploaded video is therefore crucial as a methodology to reach the source.

Overall, when it comes to visual and audiovisual material verification and previous methods did not work out, professionals have the option to look at the content. In this case it is very much as an investigator looking for small clues that could certify or deny the validity of the material. These are old methods, where for example landmarks can confirm a location, the weather can help to understand the season,

63. Claire Wardle; Craig Silverman (Edited by). 2014. *Verification Handbook, A definitive guide to verifying digital content for emergency coverage*. Chapter 3: Verifying User-Generated Content. European Journalism Centre. Maastricht, The Netherlands.

the location and the time of the day (shadows too). “Verifying the date of a piece of video can be one of the most difficult elements of verification”⁶⁴, says Wardle. To deal with that she points out to the weather cross-reference, helped by the *Wolfram Alpha* tool (which provides information to the weather of a specific location on a set date) in combination “with tweets and data from local weather forecasters, as well as other uploads from the same location on the same day.”

In the case of videos, when people are speaking, it is possible to ask an expert to analyze the accent, by which it is possible to guess the origin of the speakers.

As a last method of this surely not complete list, is pointed out by Alfred Hermida: “Applied to the discipline of verification, a new literacies approach suggests journalists adopt a more collaborative method to determining the truth that, in theory, could be reached through an iterative process played out on networks such as Twitter.”⁶⁵ Collaboration as a verification strategy can be of two kinds: among journalists and with the audience.

The collaboration between journalists is a practice still not often adopted. The reason is that as any other business, every news organization wants to be the first one to release a news and be the first to release an unverified story seems to be more important than to be second and have some evidence. However, there are signs of a collaborative method among professional journalists, who share on Twitter what they know and what they still don’t know for sure, creating a conversational space on social media where the known (and the unknown) is being spread. The Tow Center for Digital Journalism brought the collaborative approach to a further stage: they developed *Emergent*, a platform where rumors get labeled as verified or not. “Emergent is a real-time rumor tracker. It’s part of a research project [...] that focuses on how unverified information and rumor are reported in the media. It aims

64. Claire Wardle; Craig Silverman (Edited by). 2014. *Verification Handbook, A definitive guide to verifying digital content for emergency coverage*. Chapter 3: Verifying User-Generated Content. European Journalism Centre. Maastricht, The Netherlands.

65. A. Hermida. 2010. *Twittering the News*. Journalism Practice 4:3, 297-308. Routledge. London, UK.

to develop best practices for debunking misinformation.”⁶⁶

About the engagement with audience, Andy Carvin carries forward a collaborative strategy, suggesting journalists on social media should definitely adopt it as well:

“When a big story breaks, we shouldn’t just be using social media to send out the latest headlines or ask people for their feedback after the fact. We shouldn’t even stop at asking for their help when trying to cover a big story. We should be more transparent about what we know and don’t know. We should actively address rumors being circulated online. Rather than pretending they’re not circulating, or that they’re not our concern, we should tackle them head-on, challenging the public to question them, scrutinize them, understand where they might have come from, and why.”⁶⁷

66. Emergent. 2015. *About Emergent*. Tow Center for Digital Journalism, Columbia University. <http://www.emergent.info/about>

67. Andy Carvin. 2013. *#ISOJ Keynote: Can Social Media Help Us Create A More Informed Public?* <http://goo.gl/f734t0>

INTERVIEWS

To better understand the complex reality that surrounds journalists in the social media era, I've conducted few interviews with professionals who work close to it. The interviews have been conducted throughout the development of the research, some of them even in early stages, meaning that not all of the interviewed belong to the later identified target of this work. Speaking with professionals close to the journalistic world have been enlightening and helpful in guiding my research and in the choice of many needs I decided to take into account in the design process that followed. Note that since the interviews had more the form of a nice chat rather than the question-answer one, I will not report the full conversation, while I will focus on those aspects useful for this research.

January 4, 2015 **Guido Romeo** – Wired Italia

The first professional I had the opportunity to interview has been Guido Romeo, data&business editor at *Wired Italia*. His profile is definitely an interesting one for my research, since it makes use of data and digital tools to carry forward better information, working along with different associations and within a news organization company (Wired Italia).

My questions, answered by e-mail, have been mainly focused to closely understand the use made of social media and other digital tools by professionals in the news organization field.

Guido told me that the research depends on the single story, where a significant impact is given by the publishing mean (web or paper). It is clear though that some cases need longer and deeper researches that heavily influence the sources choice. However he told me the online platforms he uses the most. Archives appear to be significant category in his research process: *Archivio Ansa* (Agenzia Nazionale

Stampa Associata, Italian not-for-profit cooperative world news agency), *Reuters Archive* (for economic researches) and *Archivio Sole24Ore*. My guess is that rely on verified and trusted sources is more secure and a good starting point for further researches. Of course he makes use of *Google* too, but he uses boolean operators to narrow the search and get more specific results. This is definitely a good help to make sure to get relevant information, having the possibility to exclude pages that use certain words, or by selecting a range of time in which the pages were published. Sometimes it is useful to know if there are any trends or topics that generate more traffic than others. Guido relies on *GoogleAnalytics* and *Chartbeat*, which he defines as “an effective way to tell what is trending now and what it will be almost certainly in the upcoming hours.”

Talking about the way he gets informed on a specific case, he replied that he reads articles from other news organizations, but not only. He contacts specialists or reads academic papers or technical reports on the topic. He also makes use of Twitter, but only for very broad researches, in which knowing “who says what” might be important. When I asked how much UGC is significant in his research process, he confirms my opinion saying that sometimes it is indeed very useful. However he highlighted that “some things are still out of the Twitter ‘radar’ when I work on them” referring to those topics that are still out of the public discussion.

Since the beginning of this work, one thing has been constant throughout the research and iteration process: the tool I want to design will have its core in a visual layer as an explorative methodology for researching. Therefore, I was interested in knowing whether a professional would have found useful this approach. I asked Guido his thoughts about being able to visually explore online news, based on metadata (such as publishing date, number of views, shares and comments). Guido replied that Chartbeat and Google Analytics already highlight the most popular things and that overall the visual element, intended as a data visualization, is not that interesting for him.

Even though this answer was quite unexpected, this helped me to understand that the visual approach might not be useful for every kind of research and that there-

fore it was necessary to narrow down the target users, identifying a more specific group inside of the big ‘journalists’ container category.

I was interested in knowing how much a journalist relies on confirmed sources and how his research aims to be more investigative (with risks to get into still not confirmed news). He told me that in journalism the difference is much more between primary and secondary sources, where the primary are first-person speeches or official documents and the secondary are information from agencies and news organizations. Of course the primary sources, he says, are more secure. The investigative approach is instead on a different level and since it requires a deeper verification (called ‘investigative report’ in the journalism vocabulary) is not always feasible for an online story written in few hours.

I am not sure that social network analysis is always useful. It often describes only what happens online, which is not necessarily what happens in the real world.

“Non sono sicuro che la social network analysis sia sempre utile. Spesso descrive solo quello che succede online. Che non è necessariamente quello che succede nel mondo fisico.”

– Guido Romeo

Guido pointed out an important matter when asked about his opinion on using social media analysis for researching new sources (I highlighted that moreover journalists agree that social media have a great value on this, but somehow a very small percentage uses them in this way). He thinks it is not always useful, because “it often describes only what happens online, which is not necessarily what happens in the real world.” However, he added, there are colleagues such as Andy Carvin who used this approach with wonderful results on the Arab Spring or the war in Libya. But this social media analysis, he explains, requires really high and specific skills and even though Andy talks about an applicable “method”, I am not sure whether this is suitable for every kind of reporting.

January 29, 2015 **Jelle Kamsma, Yordi Dam, Erik Willems** – LocalFocus

LocalFocus is a data management platform that “makes it easy to get new insights from datasets and to share them with colleagues and audience.” Based in Amsterdam, I was able to meet the three members of the LocalFocus team to tell each other what we do and to get possible advices and new perspectives.

Meeting Jelle Kamsma gave me the opportunity to know the news organizations world from who actually works with them and for them. As a tool for data journalists, LocalFocus has the peculiarity of being a selling product among most of the dutch news organizations, and this has definitely similarities with what my project aims to do: helping the workflow of professionals.

It is important to keep in mind that LocalFocus is a platform where journalists can “turn datasets into stories”, as the pay off on the website claims. This is however a very different approach from the one I am intended to take, since I would like to give a way to explore the conversational space around news, rather than visualizing a pre-given official dataset.

Jelle told me their focus is to provide data on a local level, making them accessible and easy to explore. The regional section of newspapers often lacks in people, money or skills to dive into data, therefore LocalFocus makes a selection of data that are relevant and gives perspectives on why those data might generate interesting stories for a local coverage. Their target is to fill the gap between data and journalists and data and audience. Data journalism is the profession that works with data and writes stories on top of them. Data become the sources of a new journalism where being able to tell and visualize data is of primary importance. Jelle, with a data journalist background, created this platform to help news organization to solve two fundamental steps in this work: find meaningful data and visualize them. “We needed a technical component, a platform to make really easy for a journalist to filter through the data and get visual feedback so he can easily compare his region to another region or to the national average.”

What I find really interesting in this project is that each member of the team has a

journalism background. This helps them to know which are the target needs and to keep in mind journalists need a simple tool, not too complicated.

On the same platform news organizations can even upload their own dataset, share them within a group of colleagues and use the same platform to visualize them. This helps LocalFocus to be a potential data management system where journalists can store everything in one place.

LocalFocus have a main section with the list of datasets available to be used and to visualize. They upload an average of 6-7 datasets every week, about any subjects: elections, infrastructures, local, regional or national. Once a dataset is chosen, the user can filter content and an interactive visualization is already made. This can even be customized with different colors or typefaces and there is the possibility to choose among different charts (depending on the data). The tool is meant to fully integrate in the journalist workflow, therefore the result can be easily downloaded as an embeddable code to put online.

As a platform where to easily get data, LocalFocus provides both OpenData datasets and other data created by their own investigation. All datasets are then available to every news organization that has a subscription, which can create a story on that. Talking about competition among different news organizations, there are no main concerns: since datasets are about local coverage, news organizations are interested in creating stories on a regional level which will tell something different compared to another region or city.

Jelle explains me their new feature *Storyboard*, which helps journalists to create a step-by-step story with different visualizations or by highlighting different parts of the same one. This has definitely a huge impact in the way a story can be told, because it helps the audience to engage with the narrative and to play with the data, raising the understanding and the awareness.

While chatting, Jelle said something that triggered me. He said that the platform suddenly gives a visual feedback to the journalist as he opens the dataset. This reminded me that often we think about the visual layer as a candy for the audience, something to attract them and to help them understand, but it is in a way just the

last thing. While it is also true that, as everyone, journalists as well work better and faster with a visual layer in between. I therefor asked them what are their thoughts about this and how did they come up with this. Yordi, graphic designer and journalist, replied: “I worked as in infographic designer for newspapers as well and I saw that when I was working with a dataset the first thing I had when

“It’s good to see visualization not only as a tool for the audience, like a presentation tool, but also as an analysis tool.”

– Yordi Dam

it arrived it was clean and raw [...] I just put it in a super ugly barchart to see the ranking. You can do it in Excel as well and it speaks by itself when you visualize it.” Of course you can sort columns in Excel and get the same result, adds Jelle, but you don’t see it in the same way. When it’s visual is way much faster. Yordi underlines: “It’s good to see visualization not only as a tool for the audience, like a presentation tool, but also as an analysis tool [...] I guess it’s a trigger for them (the journalists, Ed.) to use the data to write a story.”

This nice conversation was particularly useful to understand that it is important to target a specific journalistic genre, since there are not only different kinds of news organizations, but different topics they want to cover and different ways they are willing to cover those. Therefore a tool to explore the conversation around news or a topic is not something for all and not something suitable for all journalists. Data journalists, for example, might be interested in such an exploration only to analyze the context around a topic they are covering, in which a dataset and its visualization are enough to tell a story.

My chat with Stijn happened via Skype, since he is currently resident in Ghent, Belgium. In the call there was also Liliana Bounegru of the Digital Methods Initiative at the University of Amsterdam (my co-supervisor), who got me in contact with Stijn.

Stijn Debrouwere has happened to be the first data journalist in Europe. Even though data journalists are not one of the target groups on which I decided to focus, I was interested in talking with Stijn because of his past experiences: he has a design centered and prototyping background and later moved to the analytics field in the US for a local newspaper and for *The Guardian*. He has worked for the Tow Center for Digital Journalism in the early stages of *NewsLynx*. NewsLynx, as the page of the website itself describes, is a project born under a movement that “has arisen around measuring the impact of journalism. This movement seeks to find alternatives to the widely used metrics of page views, time on page, and social media shares, none of which fully capture the impact of a news story. Representatives of the *Washington Post*, *The New York Times*, and *ProPublica* have all openly discussed this need.”⁶⁸ Overall, Stijn told me, the project aims to find different ways to measure the impact of news that moves away from the only page views metric. For example, he explained better, if a news article got tweeted we want to know *who* was the author of the tweet, instead of how many tweeted it, because this is relevant to answer the question “Was this story impactful?”. The project, on which Stijn has worked along with Brian Abelson and Michael Keller, “seeks to create a platform for media organizations to document the short and long-term impact of investigative news stories by combining automated metrics with flexible tools for capturing qualitative insights. ”

Being impactful is important for news organizations; that is why they are interest-

68. B. Abelson; M. Keller. April 2014. *Tow Fellows Brian Abelson, Stijn Debrouwere, and Michael Keller to Study the Impact of Journalism*. Tow Center for Digital Journalism at Columbia University. <http://goo.gl/3AF4dY>

ed in knowing what makes stories go well on social media. On the other hand, when a story is successful, it means that it was also good for the society.

Stijn told me that since getting metrics, such as article views and time spent on a specific page, is kept private by news organizations (they monitor it, but they obviously do not share this kind of information), one way they workaroud this was to get the data from the shared links on social media. All news organizations in facts have a Facebook and a Twitter page (among other platforms) that they use to spread news articles and multiply the visibility. Since those platforms leave public metrics such as likes and comments or favorites and retweets, the NewsLynx team used those as a starting point for the quantitative (how many) analysis. Plus, the same platforms give full access to know which profiles appreciated or commented the shared link, making the qualitative (who) analysis available too.

After introducing the ideas for this thesis project, Stijn pointed me to the Media Cloud Initiative, which he says it follows the medium to understand social dynamics. It uses metadata to understand the state of an issue. This project seemed to be very close to what I had in mind. Another topic touched while chatting is how important is the position, within a news website, where a story gets placed. Stijn told me there are on going projects about news organizations' websites' home pages, which look where the same story gets more or less importance depending on where it is published (top part with big picture rather than small text within a side column).

My short chat with Stijn showed me how simple data (such as metrics) can be put together to build something more valuable for news organizations in order to improve the quality of journalism.

Mark Vos is currently product manager for social publishing at *Nu.nl*, the most visited news site in the Netherlands. He previously worked there as social media editor. I got in contact with Mark thanks to Jelle Kamsma, interviewed few weeks before at LocalFocus. For his work in a newsroom and with UGC, Mark has suddenly appeared an important user to talk with, in order to get a better understanding of the journalism environment and in the relation between news and social media. I met Mark in Amsterdam, at the *Sanoma Media* headquarters, where Nu.nl's newsroom is located.

Mark explained me his job: “I develop and train editors for using social tools to enrich news or to find new news or to find pictures, videos related to news.”

Nu.nl (the dutch word ‘*nu*’ means ‘*now*’) has a long history of working with UGC, mostly gotten from social networks or own social platforms (Nu.nl has two social platforms: *NUfoto* and *NUjij*). Talking about Nufoto⁶⁹, is a platform (in use since 2007) for amateur or semi-professional photographers where it is possible to upload pictures related to news. The newsroom has therefore a wide amount of available pictures that might wants to use in their stories. In that case the author gets credited and especially for starting news photographers this a good opportunity to build their portfolio. Nujij⁷⁰ instead has more the form of a forum, where the audience can discuss about news and on top of news articles, with the opportunity to inform the news organization if there were mistakes or inaccuracies. In these cases journalists starts a verification and if it turns out the user was right, the article gets updated with the corrections and the user is being publicly thanked for his/her help. I personally find both the two platforms to be a good example of engaging with the audience. The audience is not seen as the space where to “throw” information and stories, but an opportunity to provide richer (NUfoto) and more correct (NUjij) news articles.

69. <http://www.nufoto.nl/>

70. <http://www.nujij.nl/>

Apart from these two self-own platforms Nu.nl makes a strong use of other social networks, Twitter in particular. Mark shows me how they use TweetDeck, where they filters tweets to know “everything that can go wrong in Holland, so if there is a big accident or a plane goes down or a sinking boat, then we’ll see it in this stream.” He explains me that there is no filter applied to the user field (tweets got filtered by words used only) because “we want to be really early in the process, so we are aiming for automatic responses.”

Another platform they make use of is *Monitor.live*⁷¹. This website keeps track of all road emergency calls in the Netherlands, giving a real time overview of what might happen in the traffic. As Mark puts “it is a really early stage warning that something is going on, [...] indications for us to start looking into these developing stories, which could be really tiny and 9/10 times they are really tiny, but sometimes it becomes big and in really early stage we already have some information about what’s happening.”

As Nu.nl is mainly about real time reporting, the most tools they make use of concerns real time events, where being the first to announce a news is vital and of primary importance. That is why another tool they use quite often is *Flightradar*⁷². This helps them to track problems all over the world about flights and airplanes. It is indeed possible to get alerts about specific data such as emergency calls (squawk numbers identify specific requests by the aircraft) or speed, altitude, etc.

They use also another tool developed together with the University of Amsterdam. It’s called *RTReporter*⁷³, a Twitter dashboard to monitor words used by users’ tweets. The algorithm doesn’t look for a quantity of tweets with a specific keyword, but about the gain of the use of a certain word compare to the normal use of it on that given moment. “If you take ‘fire’ for instance”, Marks explains, “it’s normal that fires happen around 6-7 am and about the same hours in the evening because lots of people are cooking and there’s a lot going on, so the word ‘fire’ will have a

71. monitor.livep2000.nl

72. Flightradar24.com

73. <http://www.rtreporter.com/>

higher threshold than, let's say, about 3 pm.” This tool takes these differences into account and if there is any mutation to this normal cycle, the dashboard will notice that something is happening. “RTreporter will not tell you what you are looking for, but it will show you what is going on in the world right now”, quoting the website. Mark reveals me that he is already working on new tools for readers and editors, for example ways to make the sharing of pictures with the newsroom even easier and faster, such as in-app camera features or by making use of existing apps such as *SnapChat* and *Whatsapp*.

“It’s a really interesting time to make news
and also to develop new ways of visualizing
and representing news.”

– Mark Vos

I then showed Mark some sketches to better explain my ideas for the project. We talked about what is useful and what might help a journalist’s workflow. As already discovered with my previous research, the verification of content is something that is of crucial importance to spot news reporters and a help in this direction would definitely be useful. The idea to check (in the background of the tool) how similar are the shared pictures to each other and automatically flag if there is an older version of it online, is something that Mark thinks to be of high help in the explorative process, since it might help to ignore those images that are published as new even though they are not. It really becomes a problem when an old pictures get published as related to going-on events, because their meaning gets modified giving birth to new discussions that makes those “new images” viral. It is very important to be fast in debunking these false news and such feature would help in it.

Continuing our talk on the verification process, Mark says that “indicators is what you are looking for. Because there is never going to be this 100% trust, you will al-

ways need an editor or someone to look at this material and give the final ‘go’ or ‘not go’, but there are a lot of algorithms or other smart things that can actually help you by pre-selecting and give again indications. This brings the discussion further ‘cause lots of journalists are afraid of automatization and social media and ‘they are gonna take our jobs’. I don’t think so, I think there is actually a lot more work to seek through now than it was before, because there is this load of information and someone needs to curate it. Of course computers and other devices can help us with that, but you will always need an editor that can say ‘this is what we need’, ‘this is true’ or not. So definitely it would be great to have actually more tools and more discriminators to see what kind of material you have.”

Our talk concludes with this: “since we use UGC since 2007 it became part of the DNA of our news corporation to incorporate the input of our readers and people that are really close to something’s happening. Again it’s an indication: is not that everything they say will be published, it’s more like a lead. [...] And still you do need professionals, journalists, photographers, because they will tell another story. I don’t think you should be afraid of this because you weren’t there when it happened, we wouldn’t have this picture. And so we will continue to build tools to make easier for people to join us and to send us material. And we will be searching from an editorial point of view especially the things you are designing to have a visual representation of complex data structures... please do!” Mark adds that there are many tools that say what is happening on social networking platforms, such as trending topics. “We need to go beyond that, to get further down into the story. ‘Who started tweeting this picture?’ and ‘from which village is this?’. That’s the kind of information we are looking for.” He finishes by saying that “It’s a really interesting time to make news and also to develop new ways of visualizing and representing news.”

OPPORTUNITIES

From the research previously done (based on papers and articles written for and about digital journalism) and from the interviews done with professionals (journalists and researchers), it resulted very clearly that we are living in a period where information concerning news events is not anymore in the hands of few journalists only. The mass, the audience, is now able to share with everybody what has been witnessed with texts, pictures or videos. This is a new opportunity, rather than an obstacle, for journalists and academics that research on a public topic or on a news story. It is about the possibility to dive into an overload of information, a crowd-sourced knowledge that is constantly being shared with everyone. The first opportunity I see is about helping those who analyze this huge amount of data to not get lost. What appears to be missing, in facts, is a systematic approach. The methodologies previously exposed in the paragraph “Research methods: investigation and verification” show a lack in the way the net is being explored. Of course the exploratory approach by definition cannot be systematized. Reducing it to a checklist or a sequence of actions will not reveal the searched answer. Explore means search, investigate unknown regions. In doing this the explorer will go hunting, trying different strategies and approaches, looking for hints and clues that might lead to a way rather than to another one. The investigative process can follow some steps that repeat every time, and those can be reached only with experience, knowledge and intuition. However it is the system used that forces the explorer to stay within certain limitations. Using a simple metaphor is like using a compass for any objectives, forgetting that the only thing a compass will do is pointing in the North direction.

Analyzing the information that spreads along the Internet on different platforms might be difficult if the available tools are separated. There is therefore need for a place where the analysis can be done in one place, no matter the provenience of the content (unless it is a choice). Even though some existing tools help in solving this issue, something seems to be forgotten. Those instruments focus only on a

granular level, leaving completely out the big picture. For who researches, both level of observation are crucial because the single piece of information is completely pointless when out of the context in which it was generated and in which it lives. In addition the use of the same platforms where the content was created does not help researchers to have a clear understanding on a bigger scale. Thinking of Twitter or Google, no matter how much we filter or specify the search query. What we will get will always be a textual list of results, each corresponding to the minimal instance: a tweet for the first, a site for the second. Making sense (retrieving and picturing the context) requires therefore to read (even roughly) as many different results as possible, looking for repetitive patterns or for differences. Considering the amount of data available, this is simply not possible, especially when the research aims to keep in consideration real time content. It is therefore clear that there is need for a tool that helps to get the whole picture, that helps to explore it and to dive into it as much and as deep as desired. What will substantially change is that the researcher will then be able to consciously decide where and what needs to be further explored.

“This proliferation of user generated content represents yet another challenge and opportunity for news organizations.”

– Paul Grabowicz

Another lack into nowadays methods is when it comes to define what is the most shared content. In facts all available tools at the moment let you sort content by popularity (n of views or shares, ...), but the same content on social media (a picture, a link, a video) can be posted and shared by different people. The tool will see each of these posts as a different content, even though the attached element is the same. For the researcher this means that he will be missing an important part of the picture if similar nodes in the network are treated as different. What might

be relevant is how similar and how different is the information that spreads in the network, looking for changes over time or different groups of users.

In the workflow of professionals that use TweetDeck, the tool is not capable of answering questions related to an overall vision. To know the connections between users, one crucial thing in deep reporting and analysis in crisis situation, this kind of tools are simply not helping. They can alert when, for instance, a user mentions another, but that implies that we are already aware of a relation between two specific users. The explorer will need to hunt for clues, such as checking whom one is following and from who is followed, a long and stressful process that requires time and patience. Helping to speed up this part of the research, by automatically check relations between users, represents a good opportunity to work on.

From the research it emerged that another issue that needs to be taken in consideration is the verification process. So, even though many tools and techniques already exist, as the 2015 “verification handbook” adequately lists, what seems to be open to be improved is a system that can automatize this process. Many tools in facts still will not give a definitive answer concerning the veracity of content, but they help to give clues, which can be seen as % of trustfulness. Being able to know the results of this automatized verification could save time for those who work under time pressure, leading them to choose what needs further investigation.

Visualization: a methodology

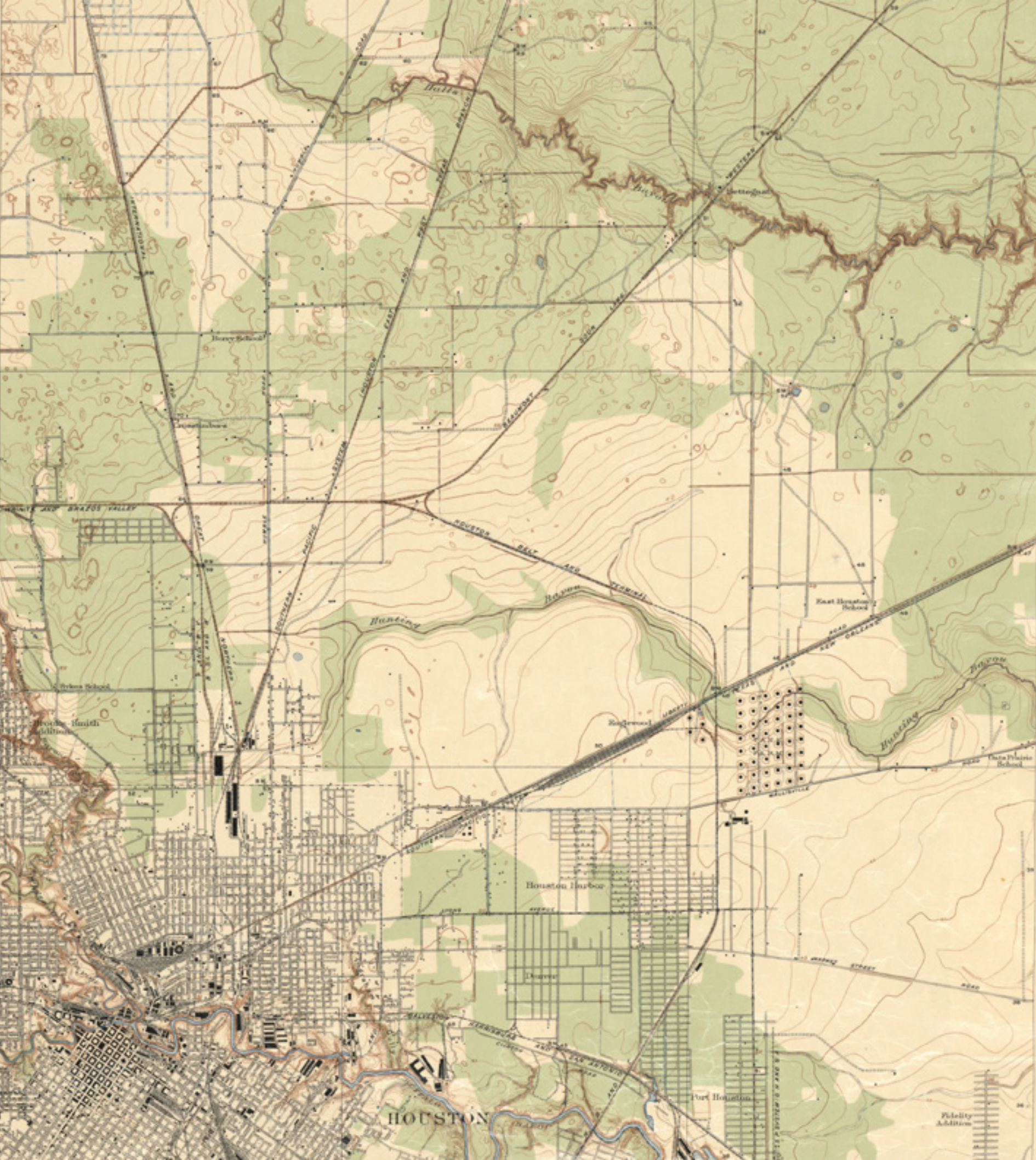
A VISUAL LAYER: SIMPLIFY, EXPLORE AND INVESTIGATE

We have seen in the previous paragraphs how difficult is for journalists and researchers to deal with such a big amount of information, where many times the single piece of content is not relevant on its own, but it brings a new value when looked in the context. The main opportunity and need is therefore the one to provide a faster way to get the whole picture, to understand which are the main correlated topics in the discussion on the social media, which the common shared content and how users interact each other.

To do so there is need for an automated methodology able to draw the big picture. As a designer, I see in this sentence the opportunity to literally translate it, by giving a very visual form: a visual layer. Why? There are several reasons why I am convinced the visualization is the optimal solution for this goal.

Lev Manovich in his 2008 “Introduction to Info-Aesthetics” says “the word ‘information’ contains within it the word ‘form’. [...] in order to be useful to us, information always has to be wrapped up in some external form.” With this statement Manovich highlights the necessity for information to be translated in a form, which can be therefore perceived with senses. For its peculiarity of being easy to access and easy to share, the visual form is definitely the one that best achieves the goal. The sight sense is also the most used one and on which it is created and consumed the most of the human communication. Hence, giving a visual form to the information makes it easier to perceive, consume and understand.

Being a visual designer myself, I cannot forget to mention the *Gestaltpsychologie* and its relevance in this field. These theories, developed by German psychologists in the 1920s, attempt to describe how people tend to organize visual elements into groups or unified wholes when certain principles are applied. Without going into the detail of all the identified principles, we can roughly summarize that the human mind seeks patterns and connections. That means that even in front of complex figures our mind is always looking for meaningfulness. Using this subconscious and complex phenomenon let the visual representation of data much faster



[13] Harris County, Texas, Settegast Quadrangle. 1922. Perry-Castañeda Library, Walter Geology Library and Dolph Briscoe Center for American History at the University of Texas.

to explore compare to the database from which it was generated. It is important to notice that the recurrence of the sight sense is extremely useful in the exploration stage of a research, which is meant to give clues, ideas and intuitions on where to dig more for a better understanding. It is more about the perception and about using a heuristic approach rather than an analytical process.

The visualization, as I intend to use it, “more than a mode of presentation, becomes, first of all, a method for exploring a particular dataset.”⁷⁴ This distinction, as highlighted by Uboldi et al., is of significant importance. Visualizing is often seen as an embellishment, something unnecessary created at the end of a process (any kind of process) and used to wrap the result in a more appealing packaging, merely focused on a ‘sell better’ goal. Even though I am aware that a design methodology is not suitable for every kind of research, I would argue that many disciplines might benefit. As Burdick points out, there are “disciplines with a social dimension for which quantitative scientific methods are proving to be inadequate.”⁷⁵ And analytic journalism and social researchers, when looking at UGC on social media, are definitely not looking for the single numbers produced. It is not about the statistics behind every post, while it is about communities, topics of discussion, relations between users and between users and content. Hence a heuristic approach, supported by a visual map, can be of better help in the exploratory process rather than an analytical approach.

I find useful to look up the etymology of the terms *investigation* and *exploration*, core values in this work. Investigation: from Latin *investigare*, from *in-* ‘into’ + *vestigare* ‘track, trace out’. Exploration: from French *explorer*, from Latin *explorare* ‘search out’, from *ex-* ‘out’ + *plorare* ‘utter a cry’. The two terms are quite close to

74. G. Uboldi; G. Caviglia; N. Coleman; S. Heymann; G. Mantegari; P. Ciuccarelli. 2013. *Knot: an interface for the study of social networks in the humanities*. Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI.

75. A. Burdick. 2009. *Design without Designers*. Keynote for a conference on the future of art and design education in the 21st century. University of Brighton, England.

“Maps don’t depict a reality — they are not mimetic devices —, but they reveal or disclose a reality.”

– Gui Bonsiepe

each other in their meaning, however they are substantially different in the methodology and goal. While investigation aims to discover the truth by tracking clues and following traces in a systematic or formal inquiry, exploration is more about searching, traveling through an unfamiliar environment in order to learn about it. The difference becomes therefore more clear: in the first case there is an order in which to research and the goal is to find objective proves, the latter is instead more close to a personal level, where experience and intuitions are part of the process. I would argue that the study on social media and conversations around a topic are closer to an exploratory approach. As a matter of fact is rarely about looking for the truth, it is rather about discovery and understanding, especially when the area in question is broad and full of isolated irrelevant small pieces.

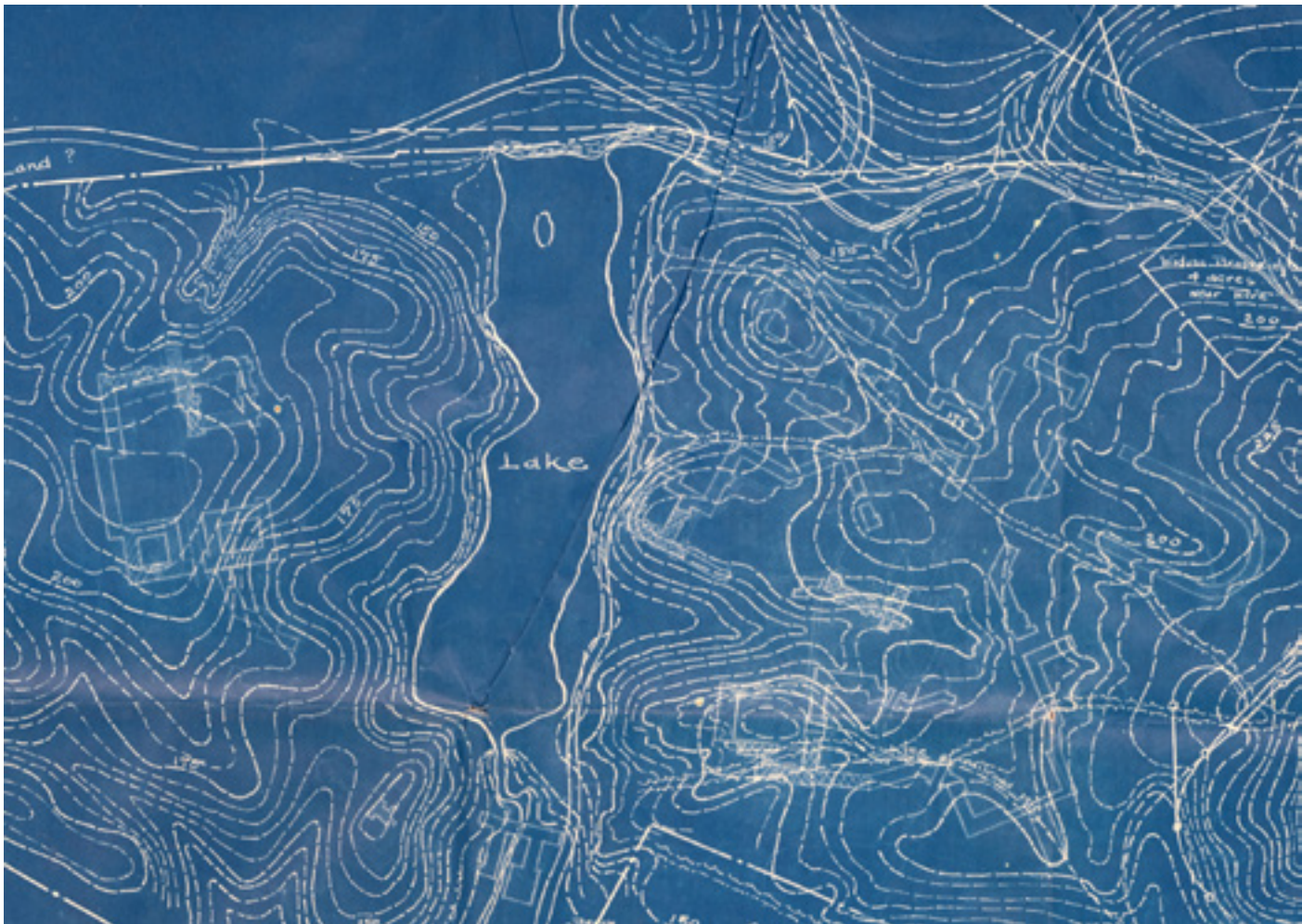
Every explorer needs a map. And when a map is not available it is necessary to create one. The use is quite self-explanatory: a map helps to not get lost (consciousness of where we are) and to get where we are headed for. Quoting the OxfordDictionaries, a map is “a diagrammatic representation of an area of land or sea showing physical features” or “a diagram or collection of data showing the spatial arrangement or distribution of something over an area.” Whether the area is physical or abstract (collection of data), the map consists of a diagrammatic representation where features are distributed over a space.

The map already works as a support in exploratory methods. For its status of being a ‘representation’ of something else, it is semiotically an interesting item. The map functions as a simplified sign (graphic representation) of a more complex object (the reality). It is only within the semiotic process of interpretation that it is possi-

ble to see the map as a *representamen* of the referent (reality).

Thinking of maps as diagrams explains why they are so useful: a diagram in fact is a “simplified drawing showing the appearance, structure, or workings of something; a schematic representation.” This definition contains few key terms that I find significant to highlight. Simplified, it means that the diagram solves a complexity issue, it reduces a complicated reality into a more simple copy of it. In this simplification process of course there are elements that get lost. In a geographic map, in example, the scale is the most relevant change, which has to lose details in order to fit in the new space. The second key term is drawing. A drawing is always a graphical sign. The drawing can be reflecting something that exists in the

[14] A portion of the campus area topographic map at the Virginia Baptist Historical Society. ~1911. VBHS Library, University of Richmond, US.



real word or something abstract that in the drawing will find its new physical form. Keeping the focus on maps, oceans' flows, temperature or elevation maps are good examples of this objectification process.

I think simplification is the very first step of the visualization process. As a matter of fact it is when facing complexity that we look at its representation to being able to understand it. Getting rid of the irrelevant noise that surrounds everything and being able to trace its most defining aspects, act as a translation of the complexity into a simplified and understandable analogue.

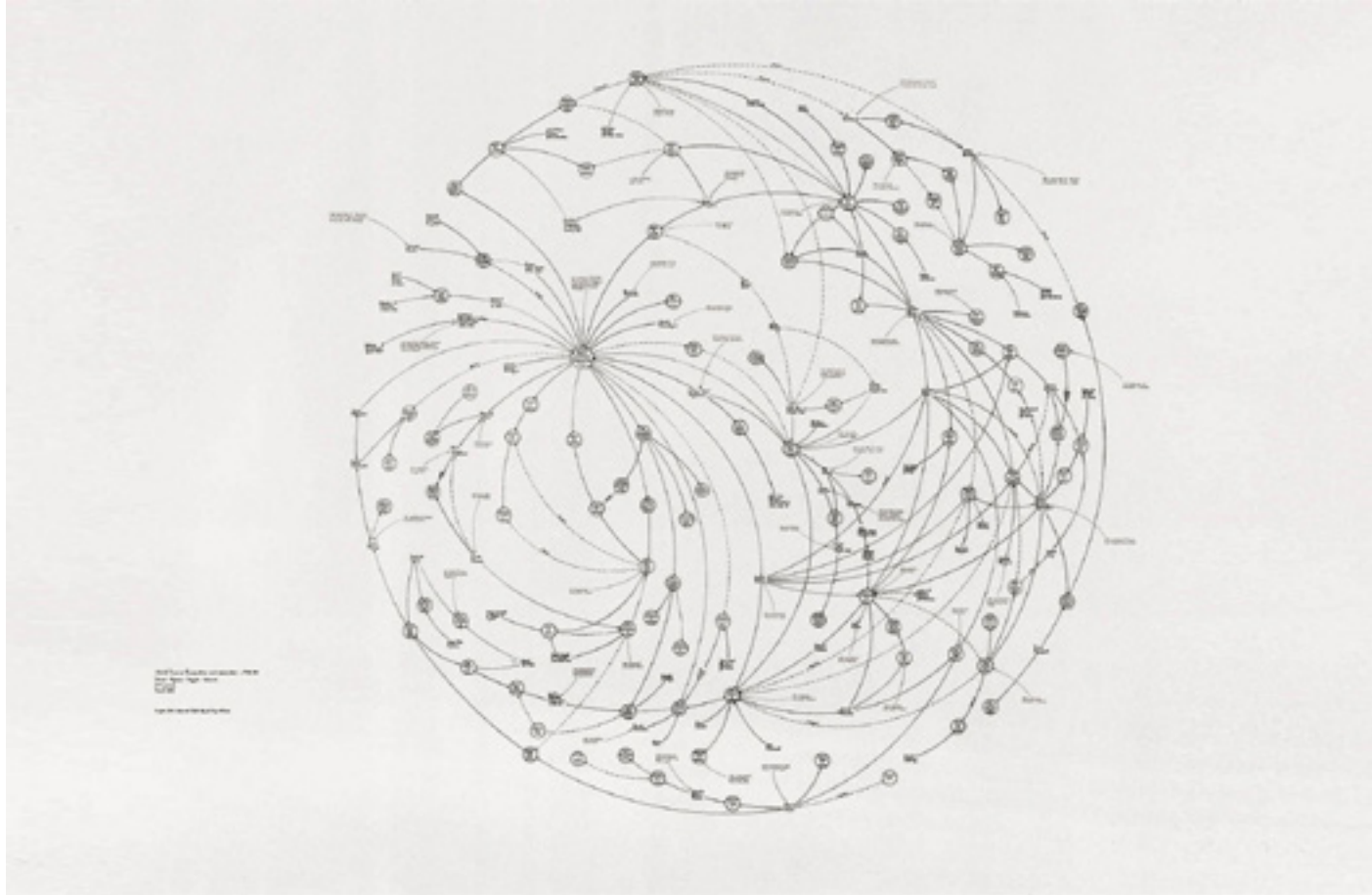
VISUALIZING INFORMATION

I find interesting to look at visualizations as maps. In one way or the other, visualizations always map something: whether it is a complex reality or a process, the goal is to represent it in order to make it clear and understandable. It is about communication. Visualizing reduces a complexity made of multiple elements (perceivable with all our senses) to a diagram that can be perceived with the sight sense only, allowing users to feel in control of it and able to dedicate different time and attention to every part of this representation. This is something that would otherwise be difficult to achieve.

Maps (strictly talking about the geo-referenced items), however, are just one of many kinds of the available visualizations of information. Those possibilities are grouped in different categories. Masud et al. help us to identify multiple approaches, all having the common goal of giving insights or in explaining a phenomenon. Those are namely: Data Visualization, Information Visualization, Scientific Visualization, Information Aesthetics, and Knowledge Visualization. It is important to keep in mind that there are no defined boundaries between the named disciplines in which visualization takes form, therefore their definitions are very similar to each other and often overlap.

I will now roughly give some indications on how to distinguish the different categories from each other. *Data Visualization* is a generic term that involves the use of graphical representation by abstracting information in a schematic form. *Information Visualization* refers instead to “the use of computer-supported, interactive, visual representations of abstract data to amplify cognition.”⁷⁶ Its purpose is not the pictures themselves, but insight (or rapid information assimilation or monitoring large amounts of data). The *Scientific Visualization* is more about the origin of the data visualized: they are physically based data in reference to space

76. S. K. Card; J. D. Mackinlay; B. Shneiderman. 1999. *Readings in information visualization: using vision to think*. Morgan Kaufmann.



[15] Mark Lombardi. 1999. World Finance Corporation and Associates, ca. 1970-84: Miami, Ajman, and Bogota-Caracas. 7th version.

coordinates. *Information Aesthetics* “forms a cross-disciplinary link between information visualization and art. It adopts more interpretive mapping techniques to augment information visualization with extrinsic meaning, or considers functional aspects in visualization art to more effectively convey meanings underlying datasets.”⁷⁷ The visual layer adopted by Information Aesthetics is therefore much more oriented to be communicative through the visual style chosen, where the perception is more significant than the correctness of the visualization. Last, *Knowledge Visualization* focuses more on transferring knowledge between people in a collaborative context. Therefore the visual representation

77. K. Friedman. 2003. *Theory construction in design research: criteria: approaches, and methods*. *Design Studies* 24:6.

is here used not for giving data insights but about getting someone to get action. Getting into the different possibilities that the visualization of information disposes, gives me as a designer the option to choose the best approach to communicate in the right way (medium and form) a specific content (message) to the chosen audience (recipient). It is therefore clear that not all the mentioned disciplines and their relative visualization approaches are equally good. Quoting Edward Tufte: “There are right ways and wrong ways to show data; there are displays that reveal the truth and displays that do not.”⁷⁸ To be more specific to our analysis, for example some of approaches work and rely deeply on the topic of the case so much that the resulting visualization is not separable anymore from the context in which it was generated. That is the case of Information Aesthetics or of Knowledge Visualization, where the visualization cannot be re-used to represent a different dataset of a different topic.

“In order to be useful to us, information always has to be wrapped up in some external form.”

– Lev Manovich

Information visualization, as part of the *analytical visualizations* category, take the data and convert it into information, in order to let the user know something and make assumptions on the data. “It’s a transformation that explores the *know-what* (or *know about*), to which we refer as declarative knowledge.”⁷⁹ I find important to highlight that in information visualizations, while visualization is the main focus, *perceptualization* (giving a perceptual form to abstract data) is the underlying goal.

78. E. Tufte. 1997. *Visual and Statistical Thinking: Displays of Evidence for Making Decisions*. Graphics Press, LLC.

79. L. Masud; F. Valsecchi; P. Ciuccarelli; D. Ricci; G. Caviglia. 2010. *From Data to Knowledge. Visualizations as Transformation Processes within the Data-Information-Knowledge Continuum*. Information Visualisation IV. 14th International Conference 2010, 445-449.

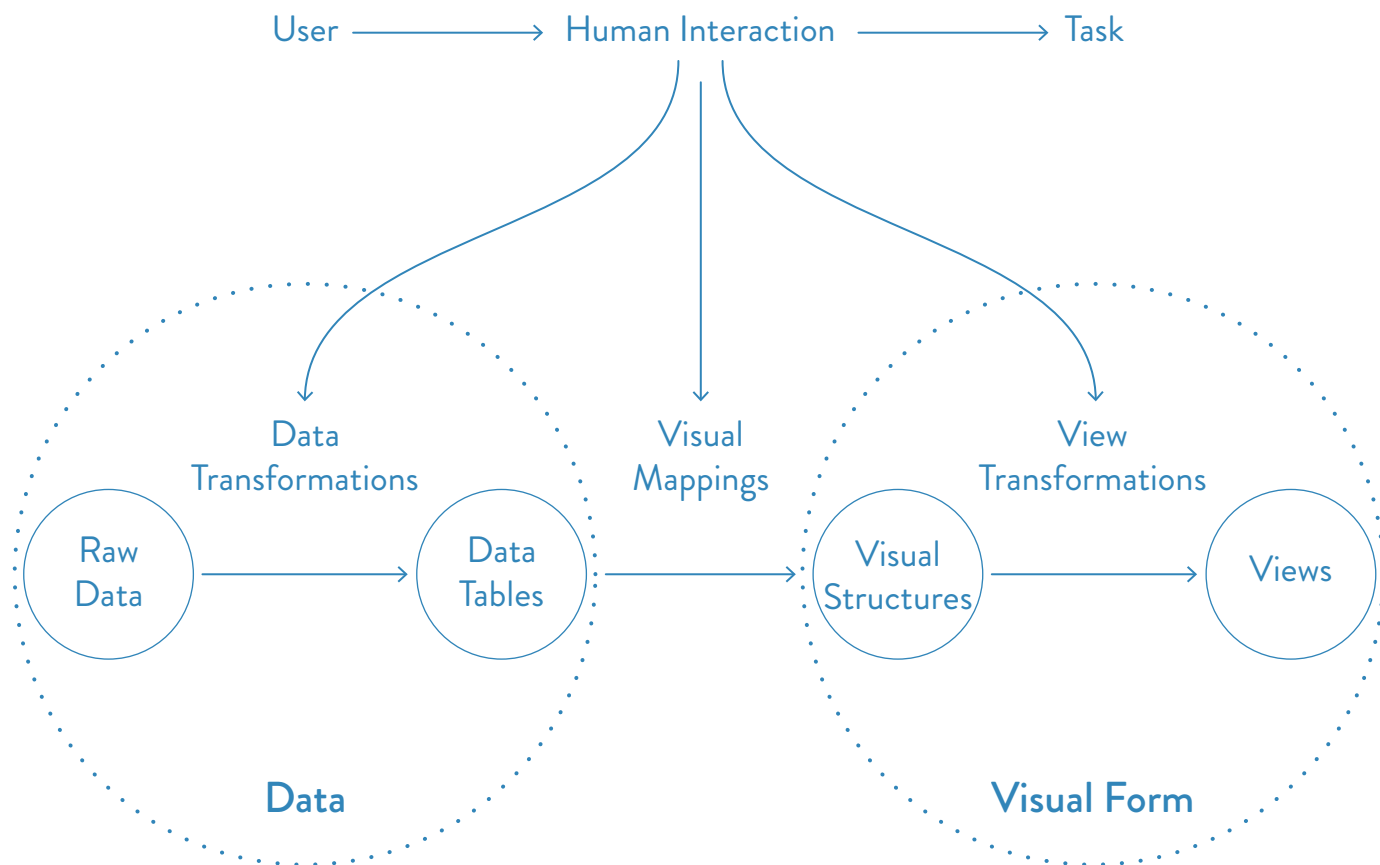
By seeing visualizations as a transformation processes, where raw data (unmeaningful per se) find a meaning by becoming information, a new question rises: How does this transformation occur?

In the previous paragraph I mentioned maps as items that support exploration. It is also true though, that “we can think of visualizations as adjustable mappings from data to visual form to the human perceiver.”⁸⁰ The process of mapping (term that derives from the map item), explains the transformation of data in their representation. Mapping, quoting the Oxford Dictionary, is an operation that associates each element of a given set (the domain) with one or more elements of a second set (the range). Even though this is a mathematical definition, we can say that since mapping the relations between data and their representation is bi-univocal, the process is (from a conceptual point of view) reversible. For this reason, in their functionalist approach analytical visualizations make the recipient able to infer data from their representation.

The data mapping process can be split in three specific steps.

The first part is the one in which raw data (abstract) get placed in *Data Tables*. Data tables are structures, called databases, in which the data have their written form (of numbers and strings of text). In this first step the transformation creates a set of relations, usually in the form of columns and rows where the first represent values and the latter represent variables. Labels are applied both to rows and columns and constitute metadata. Variables can also be of different types, namely nominal (cannot be ordered, only compared to see if identical or different), ordinal (follow an ordering) or quantitative (can perform arithmetic). During this process, called *Data Transformation*, it is typically involved a loss or gain of information. Hence, it is important to address eventual errors or missing values before the data can be visualized.

80. S. K. Card; J. D. Mackinlay; B. Shneiderman. 1999. *Readings in information visualization: using vision to think*. Morgan Kaufmann.



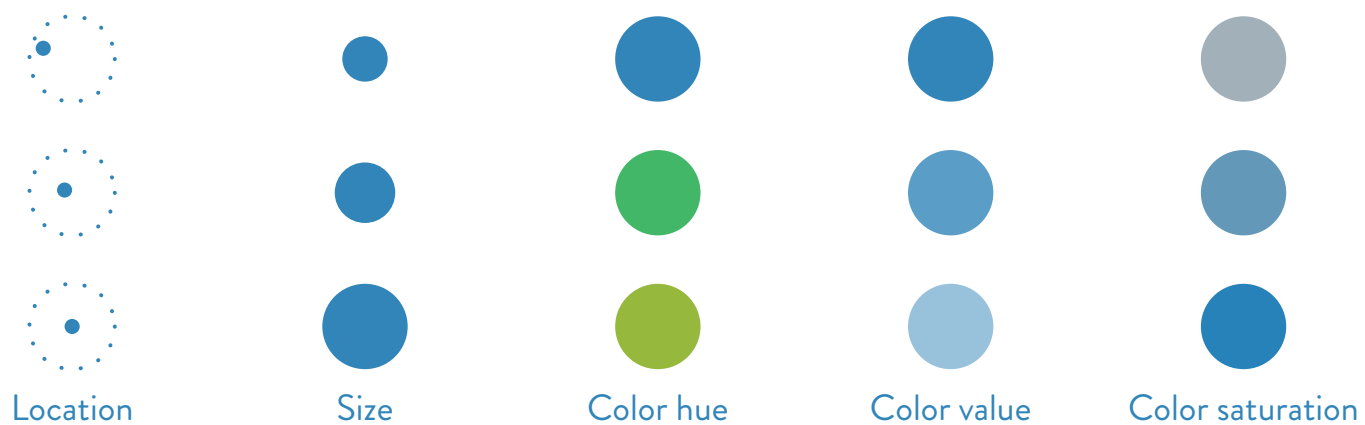
[16] Reference model for visualization. Inspired by a diagram in Card et al. 1999. Reading in information visualization: using vision to think. (redrawn)

The second part is the previously mentioned *Visual Mapping*. The transformation here occurs from data tables to visual structures “which augment a spatial substrate with marks and graphical properties to encode information”⁸¹. This is the central and most impactful step on the resulting visualization, because it is where choices on which kind of visualization and which visual forms to confer to the information will happen. It is important to highlight that the mapping process must preserve the data, maintaining the possibility to infer data from the visualization (if this doesn’t happen, we are not in front of an analytical information representation). Since there are many ways in which data tables can be ‘converted’ to visual structures, there are two parameters that need to be taken in consider-

81. Ibid.

ation in order to choose the best option available, namely *expressiveness* and *effectiveness*. A visual structure is said expressive if all and only the data in the data table are represented in the visual structure. It is in fact easy for unwanted data to appear in the visual structure, meaning that the wrong structure was chosen. The mapping must also be perceived clearly by humans. Since visualizations rely on eye perception, the designer has not to forget which visual structures are more effective in terms of perception. “A mapping is said to be more effective if it is easier to interpret, can convey more distinctions, or leads to fewer errors than some other mapping.”⁸² During the mapping process values are mapped into the spatial domain with graphics primitives and their attributes. This brings again the rules of perception listed by the *Gestaltpsychologie*. Keeping in mind these indications can help the designer to make the right choices when mapping data to visual structures, obtaining a more effective result.

Following the explanation of Card, Mackinlay and Shneiderman, other than the limits of the human perceptual system (we can’t see more than our sight allows us to see), there are also to take into account those representational limits to graphic as a medium. Bertin suggests that there is a fairly limited set of components that compose the visual structures: spatial substrate, marks and graphical properties.



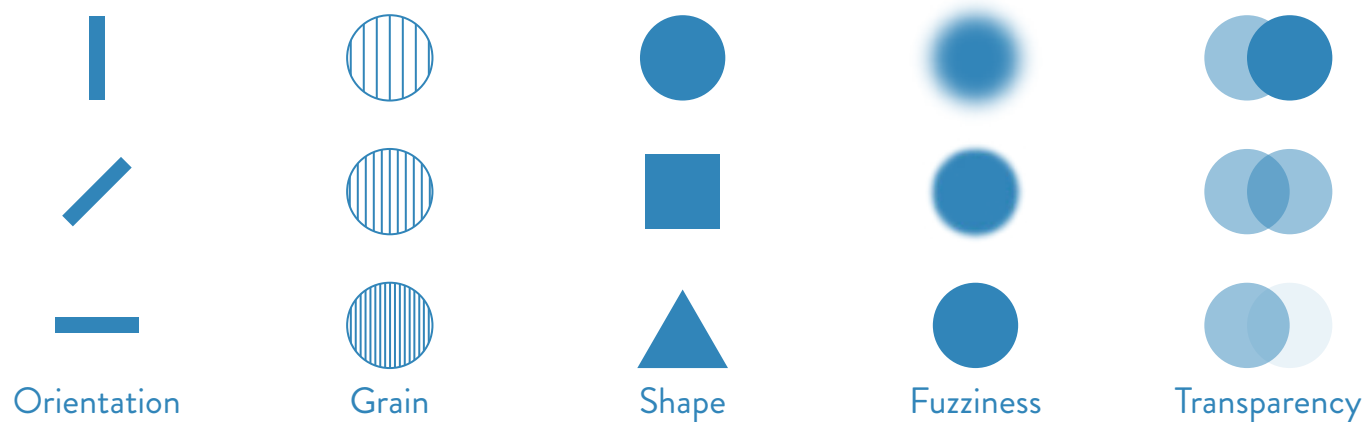
82. S. K. Card; J. D. Mackinlay; B. Shneiderman. 1999. *Readings in information visualization: using vision to think*. Morgan Kaufmann.

“The most fundamental aspect of a visual structure is its use of space.”⁸³ Space is in fact the most dominant dimension from a perceptual point of view, such that the first decision of visualization design is which variables get spatial encoding at the expenses of others. An example of the use of spatial dimension is given by the use of axes, lines that delimit a space, a canvas, in which data points are placed. To each axis is assigned a property that defines the position of the element in the bi-dimensional space (can work also for 3D visualizations).

Marks are the visible things that occur in space. They define the appearance of those data that got encoded spatially. Marks can be of four elementary types: points, lines, areas or volumes. It is interesting to notice that unlike their mathematical counterpart, points and lines in their graphical dimension, actually take up space becoming visible. Point and Line marks can be employed to represent graphs and hierarchies, allowing the representation of relations among objects, which can be connections (link between elements) or enclosures (one element contained into another).

Bertin identifies seven visual variables (also called *Retinal Properties*, since their distinction by the retina of the eye is sensitive independent of position) in which location, which coincides with the spatial substrate, occupies the first place. All the

[17] Bertin's visual variables applied to point symbol sets. (redrawn)



83. S. K. Card; J. D. Mackinlay; B. Shneiderman. 1999. *Readings in information visualization: using vision to think*. Morgan Kaufmann.

remaining ones can be associated to marks, increasing the possibility to encode more information. Those variables constitute the third and last component of Visual Structures. They are namely: size, color hue, color value, grain, orientation, and shape. Morrison and McEachren suggested the addition of four more visual variables: color saturation, fuzziness (blurred or hard edges) and transparency.

The third and final step of visualization is called *View Transformation*, and is associated with the interaction the user has with the visual structures. This process creates new views from the one initially in sight, working through an additional way. It requires the visualization to be interactive, hence it exists only in Information Visualizations supported by computer technologies.

THE ROLE OF VISUALIZATION IN THE KNOWLEDGE PROCESS

For the purposes of this work, Information Visualization is the approach for “visualizing data” that suits at best. In fact, due to its characteristics of being computer-supported and oriented to amplify cognition, is the perfect solution for representing abstract data (the complex phenomenon of online conversations) and to provide insights. In addition, it is clear how the other mentioned approaches would not work for this aim: the tool needs to rely on an interactive platform to give the freedom to explore on different levels in order to make the use richer and simpler. Due to the need of being used for different and unknown topics (the researcher will decide to explore whatever he/she wants), all the approaches that make use of particular aesthetics will not be automatically adaptable to different thematics. Last, Knowledge Visualization is oriented on the sharing of an already given knowledge, while the tool will serve as a way to expand cognition. For these reasons from now on I will focus on Information Visualization as core of this thesis.

Cognition, say Card, Mackinlay and Shneiderman, “is the acquisition or use of knowledge. This definition has the virtue of focusing as much on the purpose of visualization as the means.” They continue their definition of Information Visualization highlighting the value of insights rather than pictures and “the main goals of this insight are discovery, decision making, and explanation. Information visualization is useful to the extent that it increases our ability to perform these and other cognitive activities.”⁸⁴

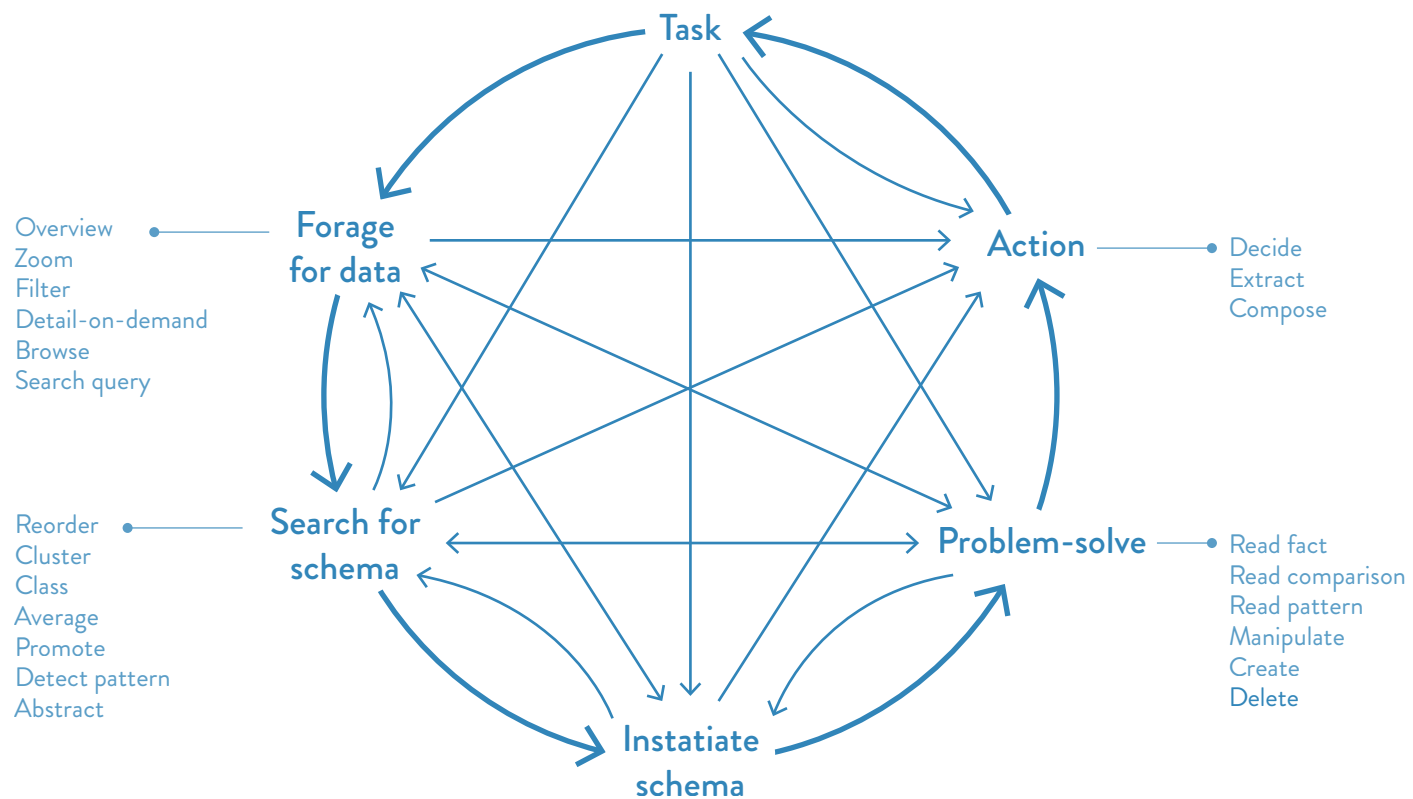
The cognition amplification at the center of Information visualization happens through the *Knowledge Crystallization* process. This task “is one in which a person gathers information for some purpose, makes sense of it by constructing a representational framework and then packages it into some form of communication or ac-

84. S. K. Card; J. D. Mackinlay; B. Shneiderman. 1999. *Readings in information visualization: using vision to think*. Morgan Kaufmann.

tion.”⁸⁵ This process is characterized by the use of large amounts of heterogeneous information, but it is guided by relatively well-defined goal requiring insight into information relative to some purpose. Some steps are identified. *Collection of data* is the first step, followed by the *search for a schema* (representational framework). Once found, the schema is applied (*instantiate schema*). Through the schema is then possible to apply the analysis and solve the goal (*problem-solving*), which is followed by the packaging of the found patterns in some *output form*.

[18] Knowledge Crystallization.
Diagram in Card et al. 1999. Reading in information visualization: using vision to think. (redrawn).

Information visualization is seeing by Card et al. as applicable to the second and third step, supporting the search for a schema and the use of the best working one. From my point of view, all the features that they put together with the in-



85. S. K. Card; J. D. Mackinlay; B. Shneiderman. 1999. *Readings in information visualization: using vision to think*. Morgan Kaufmann.

“Exactly at this point designers ought to step in, because they have — or are supposed to have — expertise in reducing cognitive complexity and help to present information by designing the interface between the information source, the data and the reader.”

– Gui Bonsiepe

formation gathering can be moved to the use of the schema itself. In fact by the use of computers, interactive visualizations give users new ways to explore information. I find this the very best feature of information visualization compared to other forms of data representation. A user can detect visual patterns and this can lead him to investigate further by searching for a specific element in the data. Therefore the *overview*, *zoom*, *filter*, *details-on-demand*, *browse* and *search query* become the basic options in the tool, in order to help the user to guide his research through the visualization itself.

Those are called *visual knowledge tools*, as they arrange information in order to reveal patterns or they allow the manipulation of information for finding patterns, or they allow visual calculations. The emphasis is on their being controls and not just presentations.

But how does visualization amplify cognition? A 1987 study by Larkin and Simon illustrates some reasons why visualization can be effective. They compared solving physics problems using diagrams versus using non-diagrammatic representations. The findings are interesting and show that diagrams helped in mainly three ways.

“(1) By grouping together information that is used together, large amounts of search were avoided. (2) By using location to group information about a single element, the need to match symbolic labels was avoided, leading reductions in search and working memory. (3) In addition, the visual representation automatically supported a large number of perceptual inferences that were extremely easy for humans.”⁸⁶

Card et al., who reported the preceding study in their 1999 *Readings in information visualization: using vision to think*, illustrates mainly six ways in which visualizations can amplify cognition: increased resources, reduced search, enhanced recognition of patterns, perceptual inference, perceptual monitoring, and manipulable medium.

As seen also in the knowledge crystallization scheme, the aim of the process is an act that follows a problem-solving. Therefore visualization, as part of the process, is seen as a mean through which to solve a problem, to achieve purposes. As Masud et al. highlights:

“Design is not pretending to provide univocal position; as designers we need to start from the purposes of communication focusing the reflection on the processes rather than the outputs. [...] From a designer’s perspective visualizations represent the process that moves from data to knowledge, where each visualization is seen as a transformation artifact within the data-information-knowledge continuum.”⁸⁷

It is in this light that visualizations can “act as communication tools.”

86. S. K. Card; J. D. Mackinlay; B. Shneiderman. 1999. *Readings in information visualization: using vision to think*. Morgan Kaufmann.

87. L. Masud; F. Valsecchi; P. Ciuccarelli; D. Ricci; G. Caviglia. 2010. *From Data to Knowledge. Visualizations as Transformation Processes within the Data-Information-Knowledge Continuum*. Information Visualisation IV. 14th International Conference 2010, 445-449.

It is important to highlight the difference between Data, Information and Knowledge that together creates the so-called DIK continuum. Raw data, as Bellinger, Castro and Mill refer, “it simply exists and has no significance beyond its existence”⁸⁸. That is that raw data do not say anything more than their presence or not: it is only through information that data gets meaning by way of relational connection. It is only by an increase of connectedness and understanding, however, that is acquired the “appropriate collection of information, such that it’s intent is to be useful.” This is knowledge and Bellinger et al. suggests it is still possible to higher up to wisdom, which “It beckons to give us understanding about which there has previously been no understanding, and in doing so, goes far beyond understanding itself.” Masud et al. say that “in this perspective visualizations are not merely final outcomes of representing data, information and knowledge. Instead they have to be conceived as transformation processes within the DIK continuum.”⁸⁹ They are therefore means by which it is possible to move from information to knowledge (and even to understanding and to wisdom, depending on the cases).

As further explained, the transformation process consists of two parts: producer’s and user’s. The first one is the act of designing, in which data and information are selected, ordered and put in relations. For their inner nature, visualizations always opt for choices on what (how much) and how to show a given data set or information. Thus “visualizations are always information in the universal domain.”

The second part of the process is instead related to the experience the user will have with the visualization. This concerns both the interaction (how the user act in front of the representation) and the experience, therefore the actions that

88. G. Bellinger; D. Castro; A. Mills. *Data, information, knowledge, and wisdom*. <http://www.systems-thinking.org/dikw/dikw.htm>

89. L. Masud; F. Valsecchi; P. Ciuccarelli; D. Ricci; G. Caviglia. 2010. *From Data to Knowledge. Visualizations as Transformation Processes within the Data-Information-Knowledge Continuum*. Information Visualisation IV. 14th International Conference 2010, 445-449.

might follow. It is only in this second and conclusive part that the information shown can become knowledge for the user. Since knowledge is something that relies upon the human capacity to understand and make use (in the sense of usefulness), the outcomes of this part are deeply subjective and are not controllable by the producer.

INTERACTIVITY

Thinking of Information Visualizations as composed of two dimensions, the first one, representation, divides data into value and structure. The second dimension, interactivity, ranges from direct manipulation to indirect manipulation. This leads to the mentioned principle by which input and output can (and I add should) reference each other. These input and output relations form the basis of a more detailed discussion of the various user actions that must be supported by visualization.

“We cannot separate the visual aspects of both data representation and graphical interface from the interaction mechanisms that help a user to browse and query the data set through its visual representation.”

– Carla Dal Sasso Freitas.

As said, interactivity is the feature that characterizes Information Visualization and differentiates it from other visualizations techniques. Interaction becomes a necessity in analytical visualizations, which have an intrinsic focus. As a matter of fact the exploratory factor would be heavily amputee without interaction possibilities.

Yi et al. points out that “while existing research in the area [of Infovis, Ed.] often focuses on representation, we highlight the overshadowed, but very important interaction component and strongly argue that it provides a way to overcome the limits of representation and augment a user’s cognition.”⁹⁰ I totally agree with

90. J. Yi; Y. Kang; J. Stasko; J. Jacko. 2007. *Toward a deeper understanding of the role of interaction in Information Visualization*. IEEE Transactions on Visualization and Computer Graphics. 13:6, 1224-1231.

this vision on Information Visualizations, which too often forget to design an interface to help in the navigation and exploration of the shown representation. For this reason I will now focus my attention to the interaction component that visualizations deserve and need.

As seen, the *View Transformation* process explained by Card et al. consists of interaction. As they say, interactivity it is the option that turns static presentations into visualizations, which exist in space-time.

I will follow the general categories of interaction techniques identified by Yi et al. These categories, as they say, are organized around a user's intent while interacting with the system rather than the interactions actually provided by the system itself. This is an important distinction that marks the centrality of the user's in the design process. I found this categorization to be the most complete one, since it was defined after a previous research on Infovis taxonomies relevant to interaction techniques.

Selection provides users with the ability to mark a data item as interesting. This option is particularly useful in front of large and complex visualizations, where the high number of data points makes impossible to keep track of the ones that the user would like to monitor or save for later. One interaction technique would be to mark those as visually different, in order to not lose them among the multitude of others data points. Most of the times select interaction is coupled with other interaction techniques to enrich user exploration and discovery.

Explore interaction answers to the 'show me something else' need. This technique is used from the well known *details-on-demand*, which consists of giving more information only after the user asks for them (probes). As Craft and Cairns say "limitations of screen real estate and visual complexity make it difficult to provide supplementary information that a data point represents, as providing in-

depth detail about all of the displayed items is impractical.”⁹¹ The key value of this interaction is that it does not (or it should not) require a change of view, allowing the additional information to be displayed within the representational context in which the data artifact is situated.

Reconfiguration works by showing a different arrangement of the information displayed in the visualization. That is that occurs a changing in the spatial arrangement of representations, providing this way alternative perspectives on the same information in view. An example is given by different ways to sort data points.

Encoding is the technique that gives the user the possibility to switch the way data are represented. In this case the Visual Structures change, modifying the perception of the data. Every change to the visual appearances (Bertin’s visual properties) works within the *encode* technique. Interactivity is essential to help users find a proper encoding scheme.

There are then interaction techniques that provide users with the ability to adjust the level of abstraction of a data representation. Those techniques fall under the *Abstract/Elaborate* category. Abstraction in this case is to be conceived as referred to the amount of details in view, meaning that its alteration spans from an overview down to the details of individual data cases. A typical example of such technique is the zooming interaction, which simply changes the scale of a representation from an overview (all in sight) to a detail view of a single data point. A key point here is that the visualization itself is not generally altered during zooming: details simply come more clearly into focus or fade away into context.

Again, another interaction possibility that static visualization could never accomplish is *Filtering*. Through filter it is possible to show something conditionally: users specify a range or condition so that only data items meeting those criteria are presented. Variants of dynamic query controls such as alphasliders, rangesliders and toggle buttons are used to filter textual, numerical and categorical data, respectively.

91. B. Craft; P. Cairns. 2005. *Beyond guidelines: what can we learn from the visual information seeking mantra?* Proceedings of the Ninth International Conference of Information Visualisation, 110-118. IEEE.

The last identified interaction technique enables users to show related items and relations, for this reasons it is called *Connect*.

In addition to those categories I feel the need to mention few other interaction techniques (identified by Shneiderman in his 1996 “*Visual Information-Seeking Mantra*”) that in my opinion do not find a place within the preceding categorization. *History* gives users the ability to easily return to a previous state in the process of exploring data. This is particularly relevant when the interaction gives the freedom to considerably change the representation. In facts, if a user makes a mistake, he should be able to easily recover the preceding state. Another important option is the *Extracting* possibility. It is simply about giving users the opportunity to export those insights the visualization helped them to find, in order to let them use those in further tasks.

In this paragraph I illustrated different techniques of human interaction in their form of direct manipulation. While a categorization of those is essential for their better understanding, it is important to remember that the different low-level interaction option given by the system overlaps each other in many ways, some of which might still need to be explored. Also, the order in which they occur is open to the user, which is able to personally explore data in the way he prefers. As Card et al. points out “human interaction with these Visual Structures and the parameters of the mapping create an information workspace for visual sense making. In real life, visual sense making usually combines these steps into complex loops. Human interaction with the information workspace reveals properties of the information that leads to new choices.”

*A supportive
tool in the
exploration
of online
conversational
spaces*

Understanding

BRIEF

For analytic journalists, deep reporters and academics who need a support in getting insights in the information overload when researching on online news content; this tool aims to be an explorative way to analyze online platforms' public conversations by the use of an interactive visual layer that provides both a macroscopic view and a granular level of information. Unlike many social media dashboards that focus on presentation it provides a working space to conduct heuristic researches.

Considering the needs of the identified target groups and the opportunities that still need to be answered, I decided to focus on specific steps of the workflow of those professionals. What I am aiming to do is design a tool that could improve those chaotic and unstructured parts of the research process.

Thinking of those that Steven identifies as the basic functions of a journalist work, "Hunter-gatherer of information; Filter; and Explainer"⁹², I see my contribution to be totally focused in the space between the first and the second one. The explainer part is in facts related to the 'storytelling' component of a journalist's work, which corresponds to the 'findings' in the case of an academic or news analyst. For this

92. S. Ross. 2005. *Teaching Computer-assisted Reporting on South India*. In Nalini Rajan (Edited by). *Practising Journalism: Values, Constrains, Implications*. Sage. New Delhi, India.

reason this function is not one on which I intend to work on, considering also that those needs concerning the final step of a journalist workflow (the act of publishing), have already been answered with many existing tools (such as *Storify*).

My tool will instead find its existing place in providing an easier way to gather in the same space the information coming from different platforms, helping to solve the hunter-gatherer function. On a second layer it will provide a visual response, a representation of this information, helping the researcher to have a ‘vision from afar’, an overall view on the content gathered. The main work will be based on a macroscopic layer that will serve as an interface to the granular information, hence giving the user the opportunity to decide where to focus deeper and further investigations. My work finds its goal in this quote from Hermida: “The need to reduce, select and filter increases as the volume of information grows, suggesting a need for information systems to aid in the representation, selection and interpretation of shared information.”⁹³

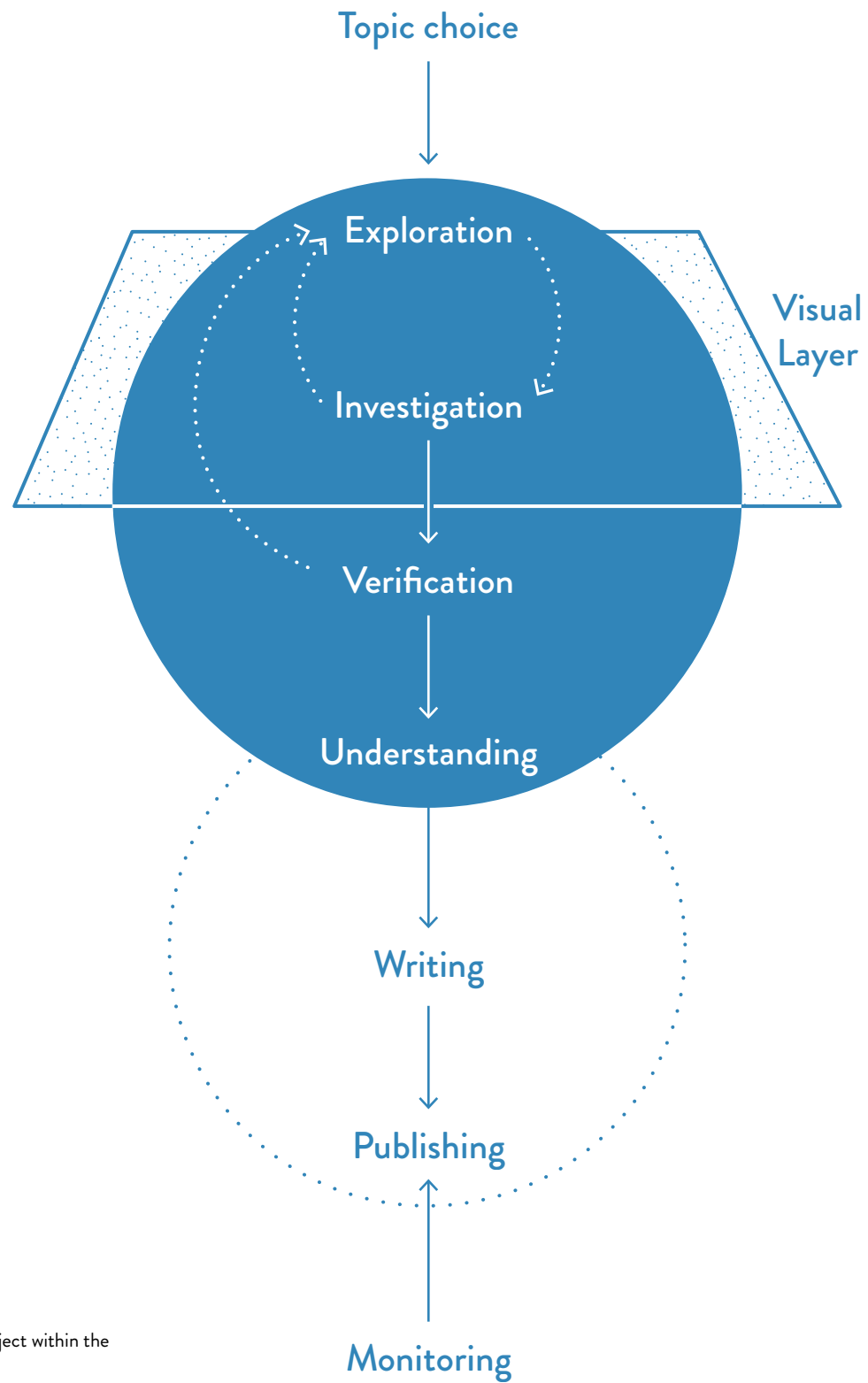
The aim of this work is not the one of replacing existing habits and methods, but to design a tool to solve the problem of list-based results, which lack in giving a bigger picture. While “all the individual pieces of information might seem insignificant, it is the communication as a whole that brings validity and significance to the users”⁹⁴, says Hedman.

To do this I will explore different ways to visualize the conversational space that gets created online around a topic or news event, keeping some aspects of the visualization open to be customized. In this way the tool will not impose itself to the researcher by providing a strict exploration rule. Instead, it will be flexible to serve the intuitions and exploratory decisions taken by the user, giving him the freedom he needs to drive the investigation as he thinks is best.

The main goal will be to create an easy-to-use instrument, with understandable visualizations. Considering the target and the use it will be done of the tool, I de-

93. A. Hermida. 2010. *Twittering the News*. Journalism Practice 4:3, 297-308. Routledge. London, UK.

94. U. Hedman. 2014. *J-Tweeters*. Digital Journalism 1-19. Routledge. London, UK.



[19] Aim of this project within the journalists' workflow.

cided to focus on the use of simple shapes that can suddenly guide to a meaning. The real opportunity here is not helping to find the right answer, but rather to ask the right question.

Throughout the thesis I often refer to conversational space and I think it is now

“The need to reduce, select and filter increases as the volume of information grows, suggesting a need for information systems to aid in the representation, selection and interpretation of shared information.”

– Alfred Hermida

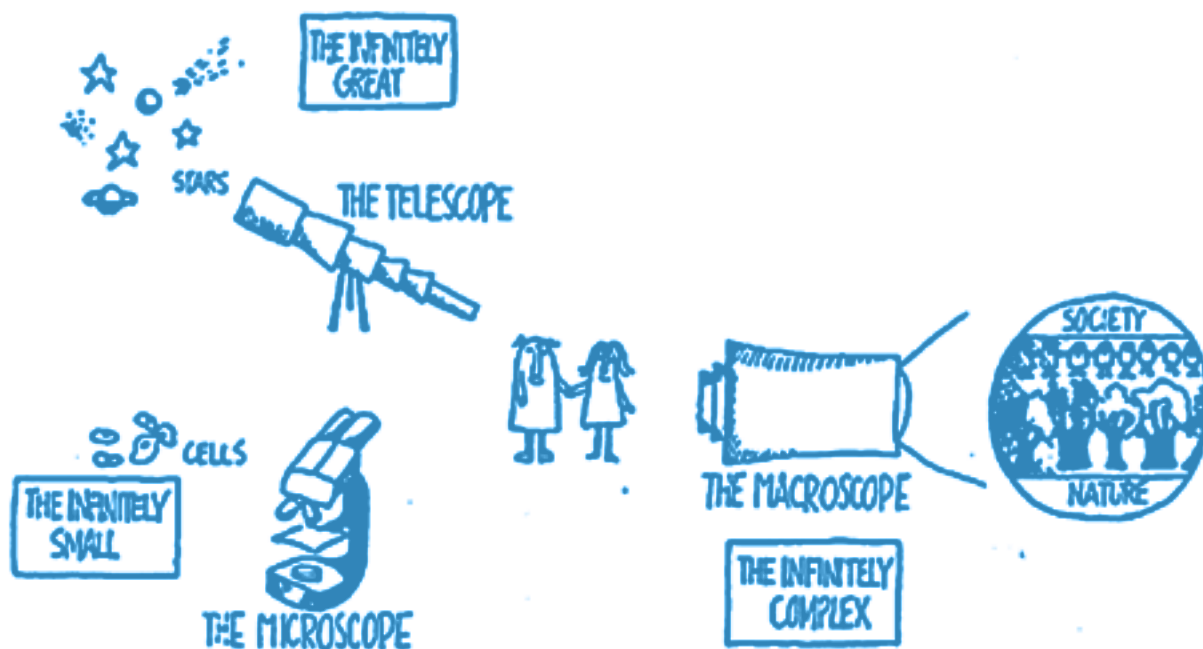
the moment to explain this term. Conversational space refers to the framework in which a conversation takes place. I refer to UGC as conversations because I believe any shared information on social media platforms is a communication act. And this act of communication is not an end in itself. As a matter of fact it is always directed to an audience, from which a response is expected (in the form of a comment, a like, or a share). It is true though that if everyone talks at the same time in a crowded arena the result is quite far from a conversation, however on social media this is possible because the voices do never overlaps, instead they stack on top of each other, creating the stream of information that we are so used to scroll. When users start to comment others' posts, a small conversation takes form, creating a small distinct space within the general one. It is also possible to notice how around a certain topic there are always other collateral ones, which extend the conversation to secondary topics of discussion. We can therefore refer to multiple conversational spaces that span across different social platforms on the World Wide Web.

FIRST THOUGHTS

Since the beginning, even before starting my research on journalism, I had the idea to support the exploration of online news content through a visual layer. Even if while researching I shifted the subject to the conversational space around a topic, the main goal did not change: to support the exploratory activity of information overload through a macroscopic visualization.

If we think about a conversational space among different entities, the picture that takes form in our mind will probably be quite unstructured and chaotic, as the voice of each participant overlaps with the others. Such a nonlinear conversation needs to be filtered and structured in order to be understood and to make the extraction of information possible. However, in the strive of leaving the pieces of content as they were created, the only structure possible is the one of the list: each part of the conversation keeps its textual form and is put in a chronological order among the others, giving to hyperlinks only the task of assolving the exploration. This is how the Twitter interface itself and how advanced tools like TweetDeck work, totally

[20] The macroscope. Illustration from J. de Rosnay. 1979. *The Macroscope: a New World Scientific System* by Harper & Row, Publishers, Inc.



leaking of an overview and generating infinite lists that might take several hours to scroll through. As a matter of fact the separation of a whole (conversation) in its compositive elements (messages) is not always good for understanding.

Quoting Scagnetti et al.:

“in 1660 Pascal says: ‘I hold it equally impossible to know the parts without knowing the whole and to know the whole without knowing the parts in detail’. This signifies the abandonment of a linear exploration and the adoption of a circular investigation that try to understand the phenomena going from the parts to the whole and vice versa.”⁹⁵

“The macroscope is unlike other tools. It is a symbolic instrument made of a number of methods and techniques borrowed from very different disciplines, [...] the symbol of a new way of seeing, understanding, and acting.”

– Joël de Rosnay

Chronology is not the only way to give an order. The Twitter interface itself, for example, provides mainly two kind of filter: by relevance (namely *Top*) or by type. After a search query is made it is possible to choose to see between *Everything*, *People*, *Photos*, *Videos*, or *News*.

Both the chronological sorting and the filtering by type options lack in some functions: an easy way to grasp what might be interesting, exploratory possibilities that move away from the browsing list way, an overview of the whole.

95. G. Scagnetti; D. Ricci; G. Baule; P. Ciuccarelli. 2007. *Reshaping communication design tools*. IASDR07 Emerging Trends in Design Research.

I would argue that all of the cited functions might find their answer in a visual representation of this conversation, which better pictures the phenomenon for its capacity to represent in a more intuitive and perceivable way dynamic relations. The macroscopic-overview approach, combined with detail-on-demand features, gives also the option the let the explorer choose where to focus more attention and so to pass from a quantitative analysis to a qualitative one.

Thinking of the identified target, I defined a set of questions that the tool should help to answer. Those questions might also be a useful test to check, at the end of the design process, whether the tool helps or not in this direction and how much time does it take to get to the same result with alternative tools or methodologies.

I imagine the tool to have a starting page where the users can see his projects. From here it will be possible to create a new project, thus a new research.

To identify the projects in a visual way I thought radarchart could be a fast and easy way to have a simple overview of the conversation around the topic inside. The value of this small representation is more aesthetic (its first goal is to provide the project a data-related thumbnail) rather than analytical. However small information could be gained from it and it would give few interesting insights especially when multiple projects are on the page: the comparison factor would mark differences. Two radarcharts could be overlapped. One referring to the geographic location of posts, divided by continents, and one that compares the ratio of posts with URLs, images and videos.

A third layer could give information about the amount of posts in the database, making it visually clear when a project is bigger than another.

A first approach for the representation of the conversational space could also rely on the way a conversation is made. For example any conversation is made of participants (who), of a message about something (what) that is expressed in some form (how). Representing those components of a conversation might help to get a better understanding, for example by showing topics of discussion in a separata visualization. However this way might lack in a general overview, which is the main goal I am aiming to achieve.

WHAT

linked/pasted images

linked/pasted videos

linked URL
↓
host + specific

related keywords
(most occurrent words)

N° of Facebook posts

tweets' audience range
(fave + rt + replies)

WHO

N° of users over time
(How many?)

users' audience range
↳ important profiles?

users' connections
(who are them?)

let the visualization work with selected users only (list)

Integration with existing methodologies
(twitter lists, trustful users)

WHERE

Location of users

geo-location of tweets

HOW

sentiment analysis
↓
geo time

Save specific things
(sidebar where to pin)

Whishlist

Alert on possibly fake profiles

Images analysis
↓
how much do they vary?

for each content look for the first time it was posted and by who

Find best match with more complete metadata (EXIF)

Search
↓
highlights in the visualization

Instead, a network visualization might be able to do that, gathering together all the different entities of a conversation in the same space. It would visually give a perceptible space to the existing but abstract space of an online conversation.

A single user can be also visualized as the central node in a hub of outgoing links. Each link can be one post, where the difference in the visual appearance could mark different types of content. The size of the outer nodes could reflect the success of the post (for example based on the number of appreciations and shares), which will be visualized also by the length of the link. By doing so a visual metaphor is applied: the more successful a post is, the farther from its origin point it went. As a matter of fact the most a post is being shared, the broader it will be its audience, reaching users not directly connected with the author of the post.

Other visualizations could focus on the analysis of geo-located posts and on the semantic analysis of their content. Being able to see those changes over time would be definitely valuable both for journalists and academics.

[21] On the left: early questions and directions of the project.

[22] Next spread pages: early sketches.



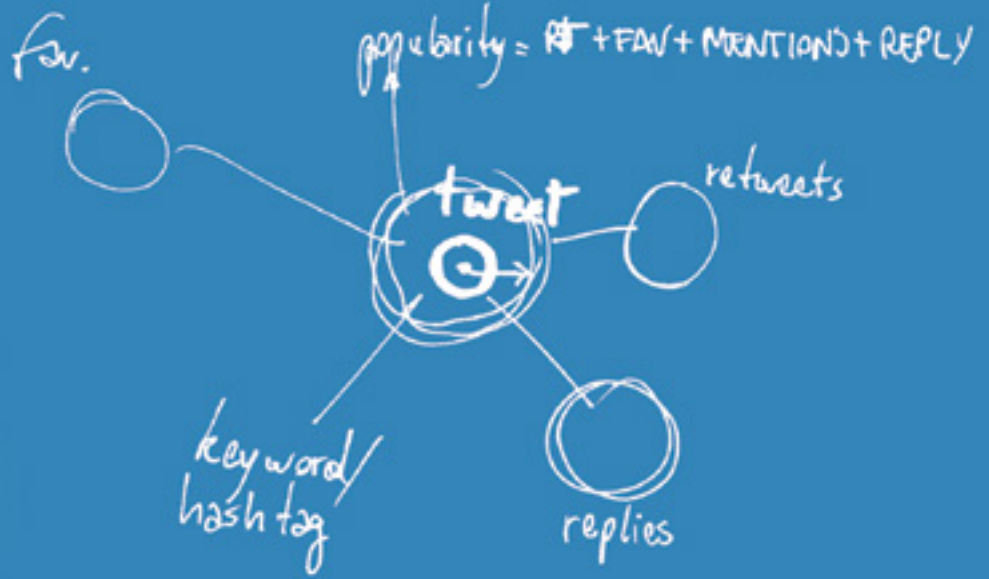
needs to be filtered
and ordered



leads
an overview

infinite!

represents better
the conversation



tweets



= mention



Overview (home)



Simplified image of a search

- where is hot
- which kind of content is more shared

TOPICS

LIST
(sorted by n. of articles/content and first article detected)



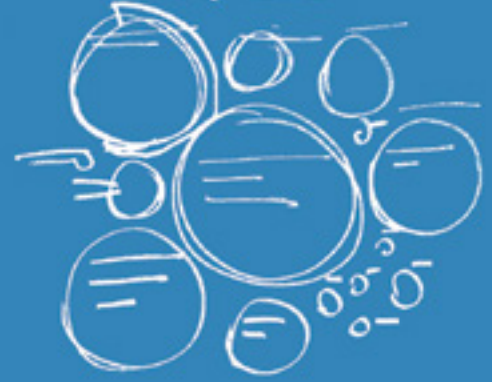
like google news

SIZE
treemap



- * Add a clue of where the topic is more trending (news websites, social media platforms)
- * Add key terms related -> most used words

BUBBLE



time

also as gallery



EXISTING TOOLS AND CASE STUDIES

It is important, for the design process, to be aware of which are the currently tools in use by the chosen target. First to know what already exists and secondly to know where there is space to innovate, where the new project will add a value. Listing and analyzing all the tools in use by researchers and journalists when facing at social media is impossible, but aiming is necessary. I will here mention and describe few of the existing tools in order to have at least a basic knowledge of the situation. In particular, even though several analytics tools exist and already work with a visual layer (usually in the format of a dashboard), I will not consider them in this analysis since they are targeted on brand monitoring and not on content value for journalism.

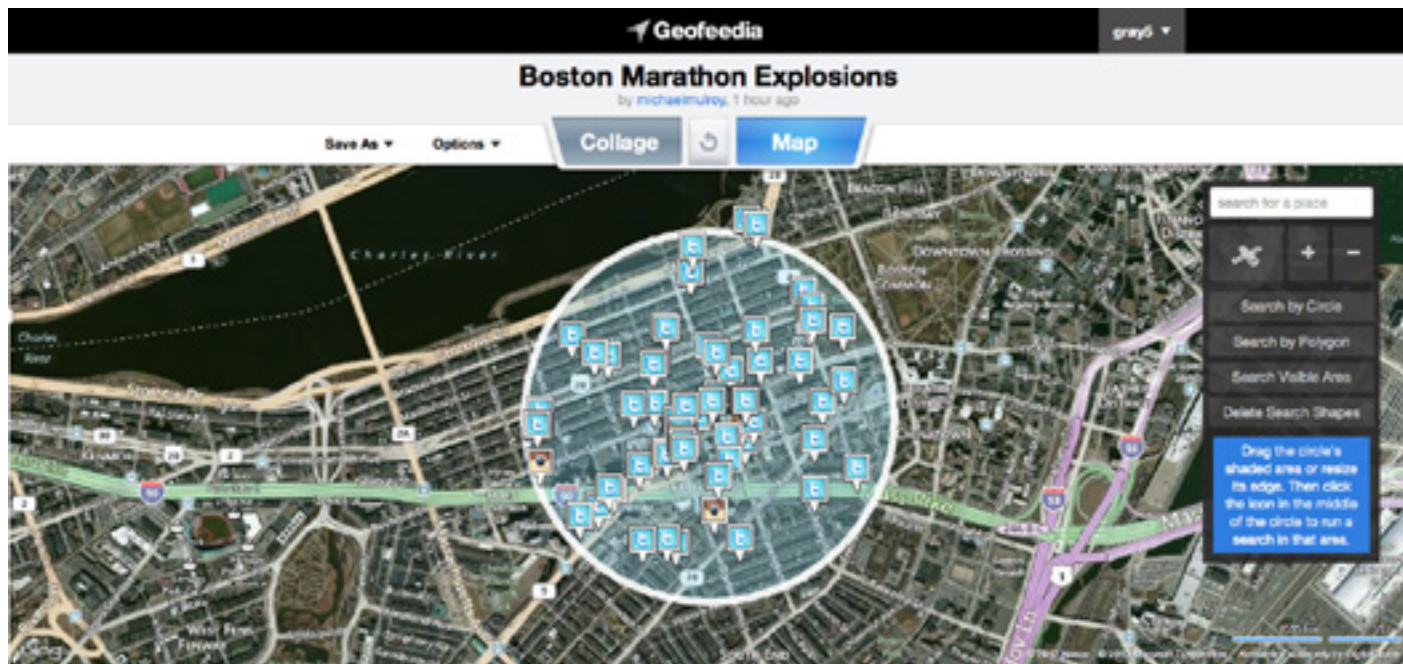
“The huge scale of social media output means that automated tools are becoming an essential part of journalism practice. There has been a proliferation of tools to help solve this problem, but none which offer a comprehensive solution.”

– Steve Schifferes

Trendsmap Solutions presents itself with this pay-off: “Real-time news-gathering and storytelling tools for broadcasters and publishers.” It is a tool that returns a visual map of geo-located content. It helps to gain detailed real-time visibility over breaking news, major events and niche topics, globally or within a region; it supports queries for any topic and boolean operators. It also lets filter by location, time, hotness and language helping to identify the most popular tweets, images, videos, links, people, locations, hashtags and words for any topic. It seems to be oriented very much into the publication of these visualizations, let-

It seems to provide a sufficient and clear analysis of Twitter activity, however it is missing some functions for a deeper and more precise analysis, such as the option to export the data. Placing on the map users and hashtags together does not seem to be the best way to gather information, in addition there is no visual difference other than the use of the *at* symbol or the *octothorpe*, making less easy the differentiation of the two elements.

[24] A screenshot of the tool Geofeedia. <http://geofeedia.com/>



The already mentioned *Geofeedia* is a tool for geo-located content discovery. It searches across Instagram, YouTube, Twitter, Facebook, Flickr, Picasa and Viddy for real-time, location-specific photos, videos and text. It provides an interactive map visualization, where content is placed with a pin marked with a symbol representing the social media where it was published.

As they put it, "breaking news doesn't wait for hashtags. Access real-time photos and videos from any location in the world, across multiple social media sources, with a single click." It is clear their help in real-time content discovery, of particular help to real-time reporters. This feature is tremendously relevant for identify

eyewitness sources and to focus on specific location for investigative purposes. The platform gives also access to a historical archive of content, helping to see old posted content in comparison to the new one. Also in this case the visualizations are embeddable to enrich any story experience.

The content is also browsable through a gallery filtered by the specific area selected on the map. The area can be set as an equal distance from a point or drawn directly on the map to suit specific needs.

The tool provides a built-in analytics view to filter content by many options. It shows few charts (barchart, linechart) related to the selected area and gives the option to export data for further analysis in other tools.

Geofeedia focuses on geo-located content only, leaving part of the discussion out. This point is particularly relevant knowing the geo-located posts are in a very low percentage; however this choice provides more secure and easily verifiable content.

Google provides professionals in the media sector a set of tools reviewed as they were designed for them. Those go under the *Google Media Tools* platform, a place where all tools are explained. It looks more as guidelines for journalists on how they get the best from available Google products.

They are not tools designed for journalists, however they are here explained in all their advanced features, which can help a journalist's work.

They span a varied range of categories: Gather and Organize, Engage, Visualize, Publish, Develop, and others. Interesting for me are the tools in the Visualize section, which are the known *Google MyMaps*, *Google FusionTables* and *Google Charts*. I am considering these among the available tools because they offer options for visualizing data for journalists. The main flow is that none of them help users to gather data about social media and therefore any help in visualizing is very much restricted to advanced capabilities of working with datasets in order to embed them into an interactive map.

In my opinion the whole set help journalists and other media professionals, but only in the story finalization, where content is put together in a form ready to be published.

“*SocialSensor* collects, processes, and aggregates big streams of social media data and multimedia to discover trends, events, influencers, and interesting media content.” SocialSensor is a project funded by the European Union and its aim is to develop software applications that collect data from social media and automatically transforms it into relevant and entertaining content for the purposes of both news and infotainment. The project is still work in progress, but an app is available for both smartphones and tablets.

The aim of this project is very similar to the goal of this thesis: to help journalists to make sense of data generated by social media platforms.

In doing so SocialSensor provides a platform on which it collects news stories. An algorithm automatically scans a news story and provides a visual feedback (green or red flag) in order to help the verification process (however, no information on how this works is provided within the app).

[25] Google Media Tools overview page. www.google.com/get/mediatools/

Google Media Tools

Home Gather and Organize Engage Visualize Publish Develop Additional Resources

Make a visual impact with Google Maps

Media covers around the world, including *Newsweek*, *Discovery* and *The Guardian*, use maps to engage readers and tell an interactive story. You can get started with tools such as Google Maps Engine, Google Fusion Tables and Google Earth.

LEARN MORE

2.50pm First explosion
Just before the finish line.
Click on the cameras

Second explosion
About 20 seconds after the first

Lennox hotel

Welcome to Google Media Tools. Consider this your starting point to tap into Google's suite of digital tools that can enhance newsgathering and exposure across television, radio, print and online.

Whether it's refining your advanced search capabilities, improving audience engagement through Google+, or learning how to visualize data using Google Maps, this website is intended to guide you through all the resources Google offers to journalists.

Note: If you're a reporter on deadline looking to get in touch with Google PR, please send an e-mail to press@google.com or visit [News from Google](#) for additional resources.

Gather and Organize

- Advanced Search
- Google Trends and Analytics

Publish

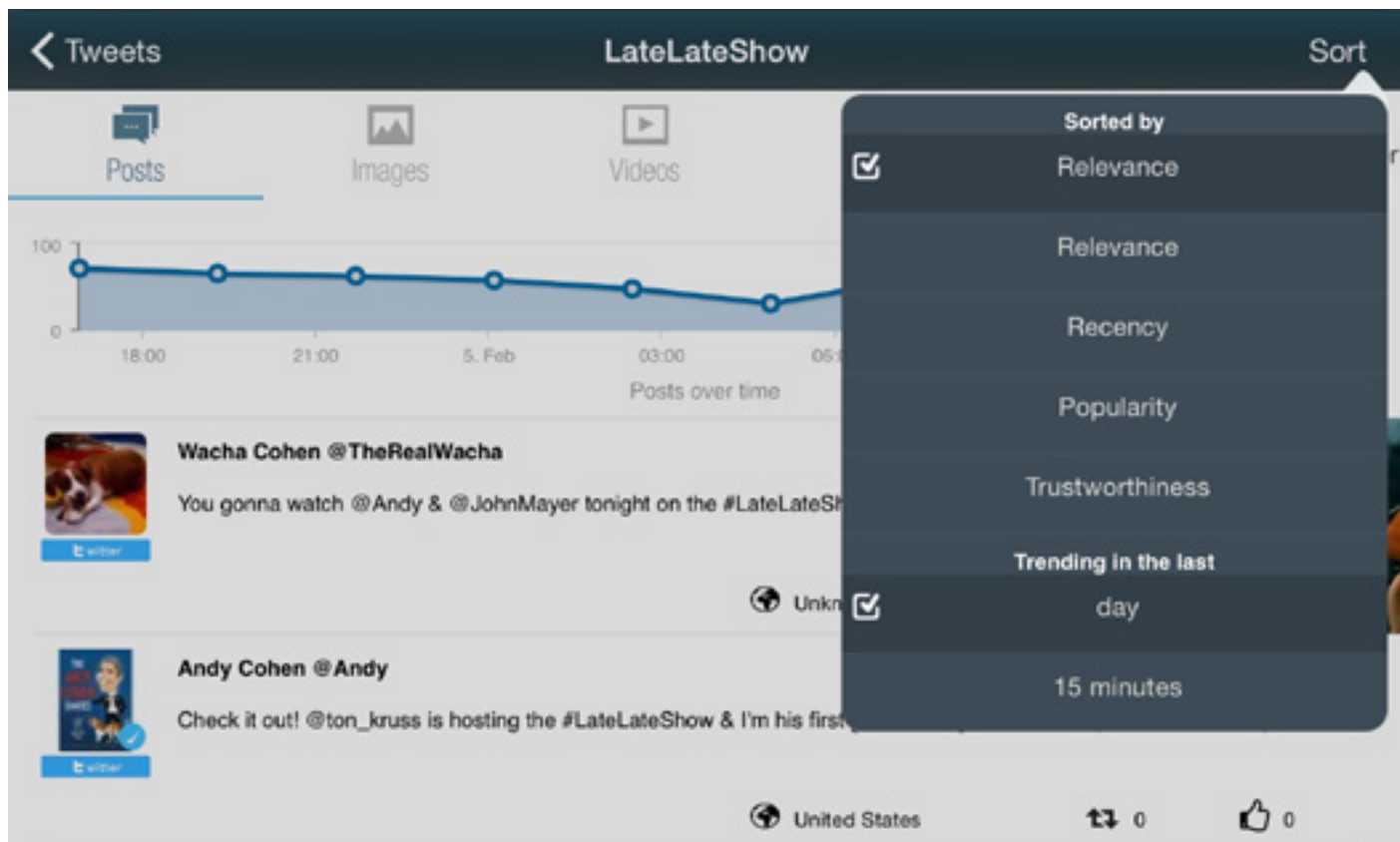
- Google News
- Google Images

Follow us on

The interface provides two main sections: a discovery list and a map. The first one is a sidebar divided in three options, namely Topics, People and Organizations. ‘Topics’ collects a list of news stories, each linking to a specific article from a news organization. If one is selected a deeper screen appears, where it is possible to browse social media content talking about that story and it is possible to browse it by posts, images, videos, headlines and hashtags. A linechart on top gives a visual clue of posts over time (but no indication about the other kind of content). The ‘People’ section contains a list of the most mentioned persons of the moment; by selecting one of them the same deeper screen as before is open but it contains posts talking about the selected entity and not only those about a certain story. ‘Organizations’ works in the same way, but recognizes the difference from a human entity.

The second option, called ‘Near me’, provides a map which shows geo-located content based on the position of the user. This might help journalists and reporters to

[26] SocialSensor screenshot of the iPad app.



be aware of what is happening around their own location. Developing a mobile app rather than the desktop one was very important for the project.

SocialSensor seems to have awesome features in the background (how data are collected and put together) and works with an interface similar to a social network. For this reason seems to be a valuable tool for real-time reporters. The tool will probably grow to support different social media platforms (right now only Twitter is shown), therefore it has a great potential.

What seems to be missing is the possibility to save findings and interesting tweets, requiring to be integrated with other research methods. Also, the overview feature still has a textual-list format, however many sorting options are available such as relevance (default), recency, popularity and trustworthiness.

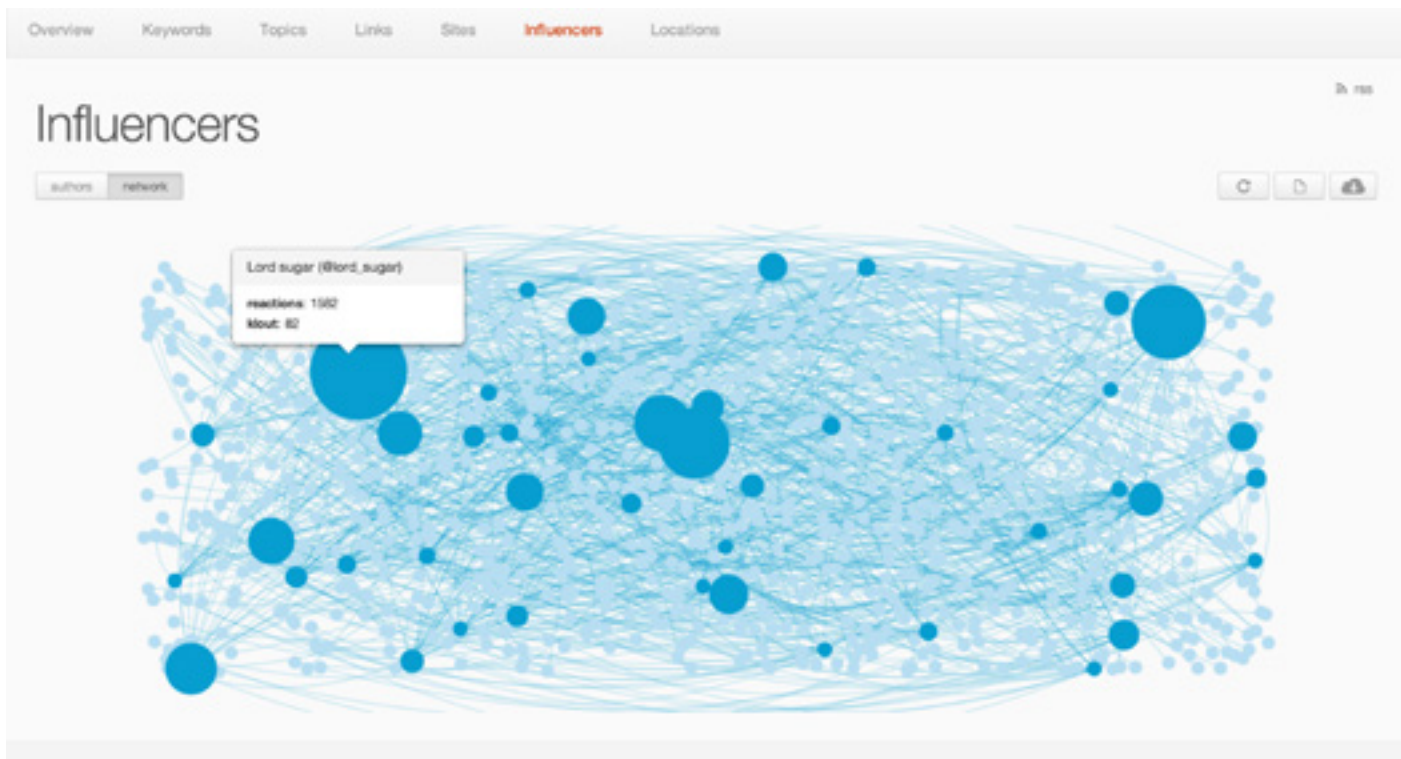
SocialSensor seems to be the only tool aiming to merge together similar content (for example it uses topics instead of hashtags). I think this is a brilliant feature to avoid repetition of similar content, however it requires high technical skills to be automated.

Pulsar is a social media monitor dashboard with particular focus on the visual layer. Pulsar provides different sub-tools to suit any need in different ways. Those products are Trac, Flow, Team, and Research. Without entering too much into the details of each, from the website is clear how many visualizations are used for getting insights.

It is interesting the use of a network visualization to understand relations between users and to spot influencers. Another useful visualization is the ‘Top Post’, which places posts as dots on a Cartesian plane where the X axes represents hours and the Y one days. The most influencing post for each hour is put in the chart and is resized by success.

I did not have the chance to try the tool in person, but it seems like it offers too much and too many ways that it makes more difficult to understand.

Again it looks like the tool is very brand monitoring oriented and does not seem to be friendly in a journalist workflow. However in the case studies provided by the

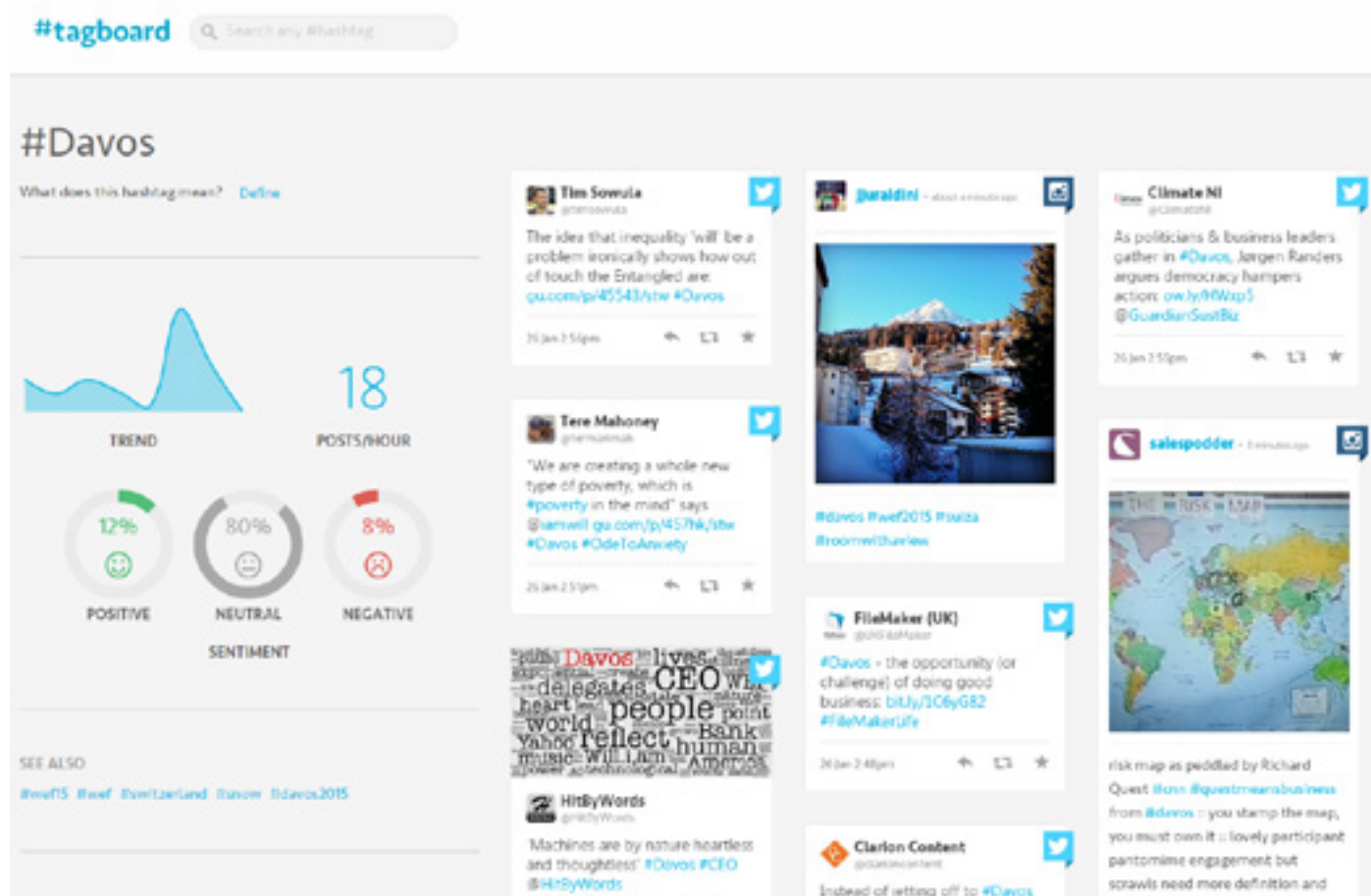


[27] Screenshot of Pulsar Platform 'Influencers' visualization. <http://www.pulsarplatform.com/>

platform there are several examples of use into the social research field, which goes in favor of its using for research matters as well.

Topicurious is a Twitter discovery tool, which helps to “filter the noise and experience social media on your terms.” Topicurious allows to find relevant and trending news, event or topic-related hashtags. The concept of the platform starts with the idea that you can search through hashtags as you do for every Google Search. To make this possible it uses a column interface where every column shows content from the selected item in the left column. It works in a similar way to the columns view in Mac OS Finder application. Every time a hashtag is selected from the hashtags list, the column on the right shows all tweets mentioning that hashtag. The list can be sorted in many ways to suits a user’s preferences.

Tagboard is a hashtag-related content browser. It lets you search for a hashtag and see all content with it. It provides content from different platforms (not Twitter



[28] Screenshot of Tagboard.
<https://tagboard.com/>

only) and it shows a small dashboard view with the trending over time, the average posts per hour, sentiment analysis summary and related hashtags.

It provides other filters options and it is possible to create a Tagboard in order to save specific searches. However it only gathers content from many social networking site and re-displays it in a different visual interface. It is good to grasp what is trending on the web, but does not really help to have an overview.

Process

DATA TRANSFORMATIONS: FROM RAW DATA TO DATA TABLES

The reference model of the visualization process, explained in the previous chapter, starts with Data Transformation, in which raw data is gathered and put together in the structured form of data tables.

In the case of this project, the difficulties regarding this first component of the process are of different concerns.

The first issue encounters when it comes to extract data from social media platforms. Those platforms and their content are publicly available to everyone (faithful to the democratic principle of the Internet), however at the same time they are actually property of the company. Facebook, for example, is very close from a sharing data point of view. And, anyway, we need to consider the amount of users who set their privacy to “share with everyone”, which is the only option that let the data to be gathered.

Twitter, on the other hand, is far more open. Through the public API it is possible to scrape new tweets (which are commonly shared with everybody, hence a much higher percentage of those are available compare to Facebook data) that match a search query (by hashtag, user mentions, etc.). For these reasons Twitter is the favorite social network for journalists for both participation in the public discussion and as a source of information.

Regardless of this technical issue, there is another one that is more conceptual: different platforms present data in different ways. For example tweets have favorites, retweets, embedded retweets as main metadata. Posts on Facebook instead feature likes and shares. We then have comments, which on the two platforms behave differently, since on Twitter a comment always need to embed a mention and that specific comment appears as a tweet itself (only a specific used grammar). These differences cannot be ignored when aiming to use the related values to be represented together with the same visual structure. However, it is possible to find the similarities and therefore apply a visual structure that works on them while keeping the differences as additional information of the single element.

The obvious similarity is the quality of the medium. In fact on the Internet, content can appear in only few forms: textual, visual, audiovisual and interactive. Since by now the interactive form, seen as the capability of content to change its displayed appearance and information due to an action by the user (such as interactive charts or website where the content *is* the interaction), is not yet integrated within social media, we can rely on the remaining three. In this way the content shared on different platform assure a 100% match. As a matter of fact, text is the form of language, which can be computationally analyzed, as well as pictures. Videos are more difficult to match, since they would require a strong and heavy calculation to be compared, however it would be technically possible. All this is actually at the base of search engines like Google, which demonstrate the feasibility of a content comparison across the Internet.

To keep the comparison level of data as high as possible, I decided to focus on social media platforms only, where data like number of posts and relations between users and content (in any form) work in the same way, avoiding strong mistakes in the creation of the data tables.

In addition, all the main social networks (namely Facebook, Twitter and Instagram) present strong similarities with each other: the use of hashtags to cluster content under a specific keyword, the use of tags (mentions) to link to a specific user on the platform, the linking of URLs or the upload of media content (pictures or video). Each of the mentioned platforms allows all of these possibilities, hence a relational structure is common and can be used as the base for transforming raw data into a data structure.

In the case of social network, even raw data follow a structure. That is because of the inner structure of the web environment, where an HTML coded page presents its content in containers nested one into each other, defining in this way hierarchies and a repetitive pattern (the same type of content is always to be found in its same context). There are basically two ways to get this kind of data. One consists in scraping a webpage that contains the desired data. In this case mathematical ex-

pressions are used to get rid of the HTML structure and to get the values (numbers or strings). The second way uses APIs, acronym of Application Program Interface. This set of routines, protocols and tools helps developers to build software applications. Therefore API allows third party apps to work with data coming from another platform. The main difference from the previous method is that it does not work with a loaded now static page, but it gets the data as they are produced. APIs usually put some limitations in the information available or in the number of records accessible (usually a certain amount of requests per range of time).

In both cases data will fill a data table, which can be in different formats (commonly in a rows/columns structure). This is the first time data are put in relations: every time there is a choice about what is a value and how to name it, data is transformed to be ready to be visually mapped.

For the sake of feasibility, in order to keep the technical and mathematical aspect of the project to those whom is competence, I decided to focus on Twitter only. The decision has been made considering the well-known openness of the platform in the research field, for its large use among journalists when they look for UGC and public discussion, and for its variety of profiles (there is no difference between users and brands as it is instead heavily marked on Facebook).

Coming up the flow I followed in order to obtain the data.

Since a tool is by definition “a device or implement used to carry out a particular function”, it is important in the design process to keep the final use, the purpose and the task that it will help a user to accomplish. Therefore it has been clear since the beginning that a research might often try to understand past conversations around a topic or event.

The method of the Twitter API to get the data demonstrated itself to be useless, since it does not let users to retrieve data older than about two weeks. To get older data from Twitter there are few services, almost all of them require an expensive subscription. However I found out *Topsy* to be very precise in the data retrieved, on top of this it is free to use.

Twitter Advanced Search

→ twitter.com/search-advanced

🔍 #MH17

📅 from 2014-07-17 to 2014-07-18



Scroll down until all results are loaded

→ GoogleChrome extension 'AutoScroll'



⬇ Save page as... → Webpage, HTML only



TextWrangler

🔍 find all lines matching 'data-tweet-id'

📄 copy results in a new document

✂ find RegEx `..data-disclosure-type="[^\"]*.*` and replace with blank

✂ find RegEx `.*data-tweet-id..` and replace with blank



✅ 53,495 tweet IDs



DMI TCAT - Twitter Capturing and Analysis Toolset

→ tcat4.digitalmethods.net/analysis/

📊 select and export data tables

[29] Gathering data flowchart.

I have chosen anyway to scrape Twitter directly to ensure the best result possible and identical to what a user would get by manually browsing through the native Twitter interface.

Thanks to an early 2014 update in the Twitter *Advanced Search* page, it is now easy to select a range of time to filter the results. I decided to conduct my research on the Malaysia Airlines Flight 17 aircraft crashed in Donetsk, Ukraine on July 17, 2014, after being shot down.

I remember, when the event happened, that on both social media and newspaper articles the keyword mostly used was MH17. For the sake of information only, I dived into the understanding of this code. MH is actually the IATA reservation code (International Air Transport Association), a two-characters code assigned to the world's airlines. 'MH' (as well as 'MAS' for the ICAO code), are assigned to Malaysia Airlines. Those codes form the first characters of a flight number, in this case flight number 17. The flight has also been marketed as KLM Flight 4103.

Somehow, the IATA code flight number became the identifier for the event. For this reason I decided to use the related hashtag as the starting point of my research.

Due to the huge amount of data produced on the topic, I had to specify a range of time from which to retrieve the tweets. Even though it would have been way more interesting to see changes over a longer span of time, I had to limit my research to the same day of the crash. As I will further explain in more detail, the cause that took me to this choice has been due to the necessity of manually scrape the tweets of interest.

To get the tweets, my query was therefore composed of two parameters only: (1) All tweets with the 'MH17' hashtag, and (2) from July 17, 2014 to July 18, 2014. Important to note is that Twitter returned only the first hour of tweets for July 18. In total I got 53,495 tweets (retweets have not been included in the results).

I decided not to specify my query any further because I want to map the conversational space regardless of the language used (which would exclude part of the conversation based only on a cultural difference) or of the place near where the tweet was generated (with this function in facts I would have included the geo-lo-

cated tweets only, which are usually an extremely low percentage). I was also not interested in filtering results based on their sentiment (positive, negative, questions). I have two reasons: firstly I do not want to filter results based on how they express a thought (the analysis, in case, could be done afterwards), and secondly I do not think sentiment analysis is good enough automatized by algorithms, which many times cannot distinguish sarcasm (which is usually a very subtle but strong way to express an opinion).

By pressing the Search button Twitter loaded the results page. I had finally access to my raw data, presented within the Twitter interface.

For the data transformations from Raw Data to Data Tables I decided to use the *TCAT* software developed within the Digital Method Initiative at the University of Amsterdam. As said on the GitHub FAQ page, “DMI-TCAT provides robust and reproducible data capture and analysis, allow easy import and export of data, interlink with existing analytics software, and guarantee methodological transparency by publishing the source code.”⁹⁶ As mentioned before, the Twitter API does not allow the retrieve of old tweets, therefore the TCAT tool could not get the data I needed. However to get all the available information of a tweet, the software needs only the tweet ID, an identificational string that univocally name an entity. The tweet ID is embedded as a metadata in the HTML structure of Twitter any time the related tweet is shown, meaning that all the IDs I needed were hidden in the code of the Twitter results page. Usually a web scraper software would do the job, but Twitter website makes use of the ‘scroll bottom to load more’ feature, which makes almost impossible to automatically scrape the HTML page. To workaroud this issue I used a Google Chrome extension called *Auto Scroll* (developed by dekuyou) to automatize the scrolling down. After around 30 hours of non-stop scrolling, the web page with the results reached its bottom, with all the tweets. This part of the process by itself shows the most severe limit of social media content exploration

96. TCAT. Digital Method Initiative. GitHub, FAQ page. <http://goo.gl/lyC2EQ>

by using traditional methods: the information overload is so big that is impossible to browse through it, even if the human eye was as fast as an auto-scrolling script. I downloaded the whole code using the Google Chrome's *File > Save Page As...* command, and selecting the *Webpage, HTML Only* option from the *Format* dropdown. Tweets are coded as `` nested inside of an `` container. After deleting everything's before and after the `` list, I filtered out all the unnecessary parts of the code. To do so I used *regex* (Regular Expressions) within the text editor software *TextWrangler*, after a couple of operations I had the complete list of 53,495 tweet IDs.

[30] Twitter Advanced Search page with the query for gathering the data.

The image shows a screenshot of the Twitter Advanced Search interface. The browser's address bar displays the URL `https://twitter.com/search-advanced`. The page title is "Advanced Search". The interface is organized into several sections for filtering search results:

- Words:** Includes options for "All of these words", "This exact phrase", "Any of these words", "None of these words", "These hashtags" (with the example "#M-H17"), and "Written in" (with a dropdown menu set to "Any Language").
- People:** Includes options for "From these accounts", "To these accounts", and "Monitoring these accounts".
- Places:** Includes "Near this place" and a link to "Add location".
- Dates:** Includes "From this date" with a date range from "2014-07-17" to "2014-07-18".
- Other:** Includes radio buttons for "Positive", "Negative", "Question", and "Include retweets".

A blue "Search" button is located at the bottom left of the form area.

Once the list of tweet IDs was successfully imported in TCAT, the software performed the complete analysis, producing a several number of data tables ready to be visualized. Some of those are in *.csv* (Comma Separated Values) format, while others are in *.gefx* (Graph Exchange XML), which are meant to be visualized as network graphs with dedicated software.

By being an analytical tool itself, TCAT is composed of three sections: data selection, overview and export. The first section serves as a filter for the second and

[31] DMI-TCAT interface open on the #MH17 data analysis.

The screenshot displays the DMI-TCAT web interface. The browser address bar shows the URL: `tcat4.digitalmethods.net/analysis/index.php?dataset=mh17&query=&url_query=&geo_query=&exclude=&from_u...`. The page title is "DMI Twitter Capturing and Analysis Toolset (DMI-TCAT)".

Data selection section:

- Select the dataset:** A dropdown menu shows "mh17 --- 53.356 tweets from 2014-07-17 16:11:40 to 2014-07-18 00:59:59". To the right, it says "447.730.233 tweets archived so far (and counting)".
- Select parameters:** A list of input fields for filtering:
 - Query: (empty) (empty: containing any text*)
 - Exclude: (empty) (empty: exclude nothing*)
 - From user: (empty) (empty: from any user*)
 - From twitter client: (empty) (empty: from any client*)
 - (Part of) URL: (empty) (empty: any or all URLs*)
 - GEO bounding polygon: (empty) (POLYGON in [WKT](#) format)
 - Startdate: 2014-07-17 (YYYY-MM-DD or YYYY-MM-DD HH:MM:SS)
 - Enddate: 2014-07-18 (YYYY-MM-DD or YYYY-MM-DD HH:MM:SS)
- An "update overview" button is located below the parameters.
- A note at the bottom states: "* You can also do AND or OR queries, although you cannot mix AND and OR in the same query."

Overview of your selection section:

- Dataset: mh17 ()
- Search query:
- Comments:
- Exclude:
- From user:
- From twitter client:
- (Part of) URL:
- GEO polygon:
- Startdate: 2014-07-17
- Enddate: 2014-07-18
- Number of tweets: 53.356
- Number of distinct users: 42.299

A pie chart on the right shows the distribution of tweets:

- Blue slice: 22.8% (Tweets containing links)
- Red slice: 77.2% (Tweets containing no links)

the third ones, where is possible to define specific queries. The overview section already shows some visualizations and a summary of the content in the database: a piechart shows the percentage of Tweets containing links vs. Tweets containing no links, while a linechart represents the amount of users, tweets, locations and geo-coded content. Both charts present a hover interaction that highlights the hovered element and displays more specific data in a tooltip. The third section is focused on providing researchers with advanced analyses of their Twitter datasets, which they can export directly.

TCAT performs different statistical analysis on tweets, users, hashtags, URLs, words, and media, creating in this way new data structures and a first hint of information. Another important analysis done by the tool is about relations between data. The data structures are transformed again by retrieving information on how many times certain relations occur together, setting up the base for a network visualization. The available networks are between users by mentions or by replies to tweets, between hashtags, between hashtags and users, hashtags and mentions, users and URLs, hashtags and URLs and few others.

The tool is very powerful and can help researchers (with no knowledge of coding or math) to have ready-to-use data tables automatically created with an excellent background work done on the relations between data. However few notes are to keep in mind: the tool works at best for who knows how to deal with data and is familiar with databases. An average user would probably get lost in the interface and would not understand the difference between many options (which are quite complete, but are missing a hierarchy in their visual disposition on the page). Since my tool will heavily depends on analyses done by an existing platform, someone could point out that my project does not add much rather than a different visual interface. However I would argue that my contribution does not primarily focus on the interface, but on simplifying the whole process by giving filtered and easy-to-explore visualizations, minimizing this way the necessity of dealing with databases.

THE VISUAL MAPPINGS LOOP

The Visual Mapping is the most important process of the visualization model: it is the one in which data tables get translated into Visual Structures, giving meaning to data and making relations and information visual, thus perceptible. As seen both in the diagram of the process and in my own experience, this part of the process deeply consists of looping tasks, in which data tables are constantly mapped and visually structured again and again, every time starting from the results obtained by the previous structure.

This looping should not surprise since it is a key element in any design process: build to dismantle and rebuild better with more knowledge. In this section I will therefore highlight the most relevant steps of this nonlinear process, during which intuitions and visual feedback guided the choice making and more work on data tables.

As mentioned in the “First thoughts” paragraph in the previews chapter, one of the first ideas consisted in creating a network. The idea has been guided by the desire to answer to this question: Who says what, to whom, in which form and when? It actually combines multiple questions, where ‘what’ refers to the actual meaning of the message (the signified message itself) and the form is the chosen signifier of the message (textual, with attached media or linked URLs). Other information that might be available can always be considered an expansion of one of the previous ones. Where a user comes from, for example, is additional information to ‘who’.

In the beginning I thought to split and filter the data table in order to focus the network on one entity type per time. For example by creating a network of relations between hashtags, or between users.

Even if this approach is correct in aiming to clearly represent those relations, I understood it was missing the ‘bigger picture’ factor. I decided therefore not to limit the network to one entity at the time, and instead to try to put more entities together as a way to portrait the whole conversation as seen from afar.

The idea was to create a bi-partite network graph in which there are multiple nodes

categories. Those nodes, entities, could be for example users, hashtags and the content shared, allowing links to describe how strong the relation is between two entities of any kind. The network would still be bi-partite because hashtags and any other content can be seen as on the same level: they do all appear as parts of the message, where the only different entity is its author. Also, in the hypothetical idea to map relations between users, those would make sense only from user to user (with a directed edge) where instead the other edges can be of two types only: dependency (of content from author) and co-occurrence (between contents). For this reason I have decided to exclude users relations among each other out of the big network, since the edges would have worked with different meanings, making the network structure in incorrect ways and therefore leading to sure misinterpretations.

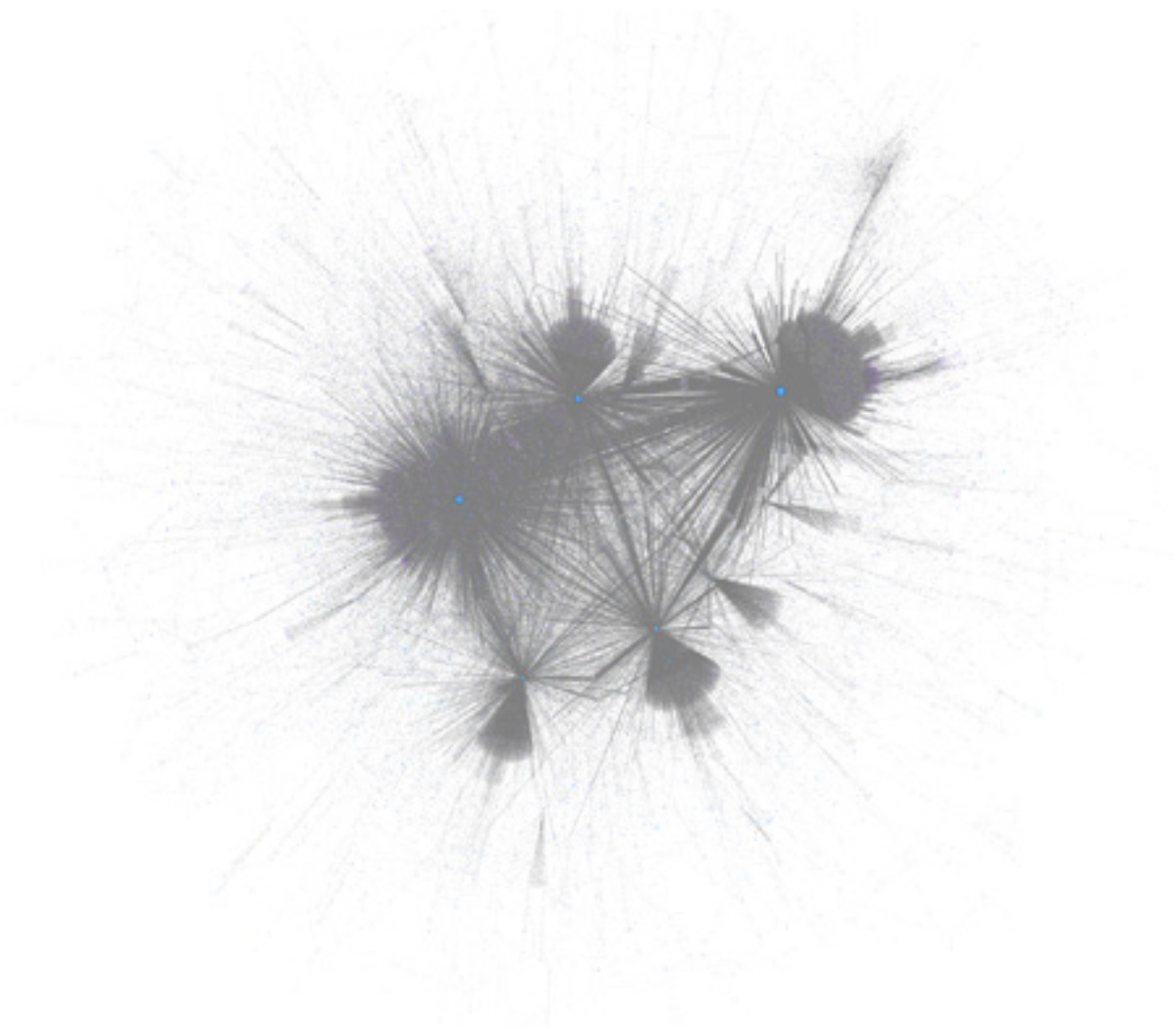
“Any serious visualization of a sufficiently complex topic should always aim exposing the complexity, the inner contradictions, the manifold nature of the underlying phenomenon.”

– Moritz Stefaner

In order to obtain the big network visualization I started from data extracted through TCAT. The analytic software already creates different network files ready to be exported. Those built the basis for the single and bipartite networks.

A very important component of this process consists in filtering. Having too many information within the visualization can create chaos and a cluttered view, totally uninviting because of its vastness, which puts severe limitations in the pattern recognition.

Starting from the hashtags frequency dataset, I filtered them by frequency, keeping only those used at least 10 times. The selection here is already strong for the



[32] Early network created in Gephi with data on #MH17. No filter applied.

relatively small database I am working on; in researches of a bigger scale however this threshold should be increased. The remaining ones were used as guide to filter relations with other entities: from the URL-hashtags bipartite network graph file I kept only those URLs that were linked to the hashtags previously selected. The same has been applied to the user-hashtags and hashtag-hashtags network files. Also in this latter process a filter by frequency has been applied, in order leave out, for example, those users who used one of the most recurrent hashtags but a very limited amount of time (a minimum of 5 was my threshold).

The new data structure just described would be enough to have a bipartite network of all the entities, however it would be based on hashtags frequency only. Thus, I

repeated the same pattern starting from the selection of the most relevant users (based on number of tweets in the database) and from the selection of the most shared URLs. At this point I had three networks, each with a main entity type with a value like the number of occurrences within the dataset (number of tweets connected to that entity). The selection of nodes occurred within the visualization software I used to map the networks: *Gephi*. This software make it relatively easy to visualize network files based on a nodes and an edges lists. It contains a database view, from which it is possible to delete or add nodes. Even though the nodes were already highly filtered, the first visualizations were still very busy and confusing. A first method in trying to enhance the pattern recognition has been to apply the visual variable of size in relation to the frequency of appearance. This resized each node, but because of the #MH17 presence, which was the term by which all tweets were gathered, the proportions were way too distant, causing the MH17 node to be huge and making all the other nodes extremely small, with difficulties in seeing further differences. I decided to leave out the #MH17 hashtag from the visualization. This might be a wrong choice, but I believe it is a good way to go. As a matter of facts as a researcher I am not interested in the pure use of the hashtag only, but in navigating its surroundings, trying to map which are the connected topics. Therefore to see the MH17 hashtag so big in the visualization is not that relevant and from an algorithm point of view it asks all the nodes to stay attached to that central

“When you interact with folks face to face, you develop a picture of the network in your mind: who else is around, who talks to whom, etc. That is not usually available online. In cyberspace, the context of the social space you’re embedded in is missing. A map can help orient you.”

– Valdis Krebs

one, which makes less easy to find distances and grouping around side hashtags. This choice created a much more clear map to navigate, where nodes finally had a visual difference in their size.

I exported the data of each of the three networks and work on them in *Microsoft Excel*. Here I verified that all the three data tables had the same structure with the same columns names and I also normalized the values of the appearance. Even though might be mathematically incorrect, this helps in the visual recognition of the most important entities. In facts if values were not normalized, only hashtags nodes would pop out (for their obviously higher frequency value compared to URLs, in example). This can help to see the most important entities of each type and it is not intended to be mathematically analyzed. It is mainly about perception and guidance.

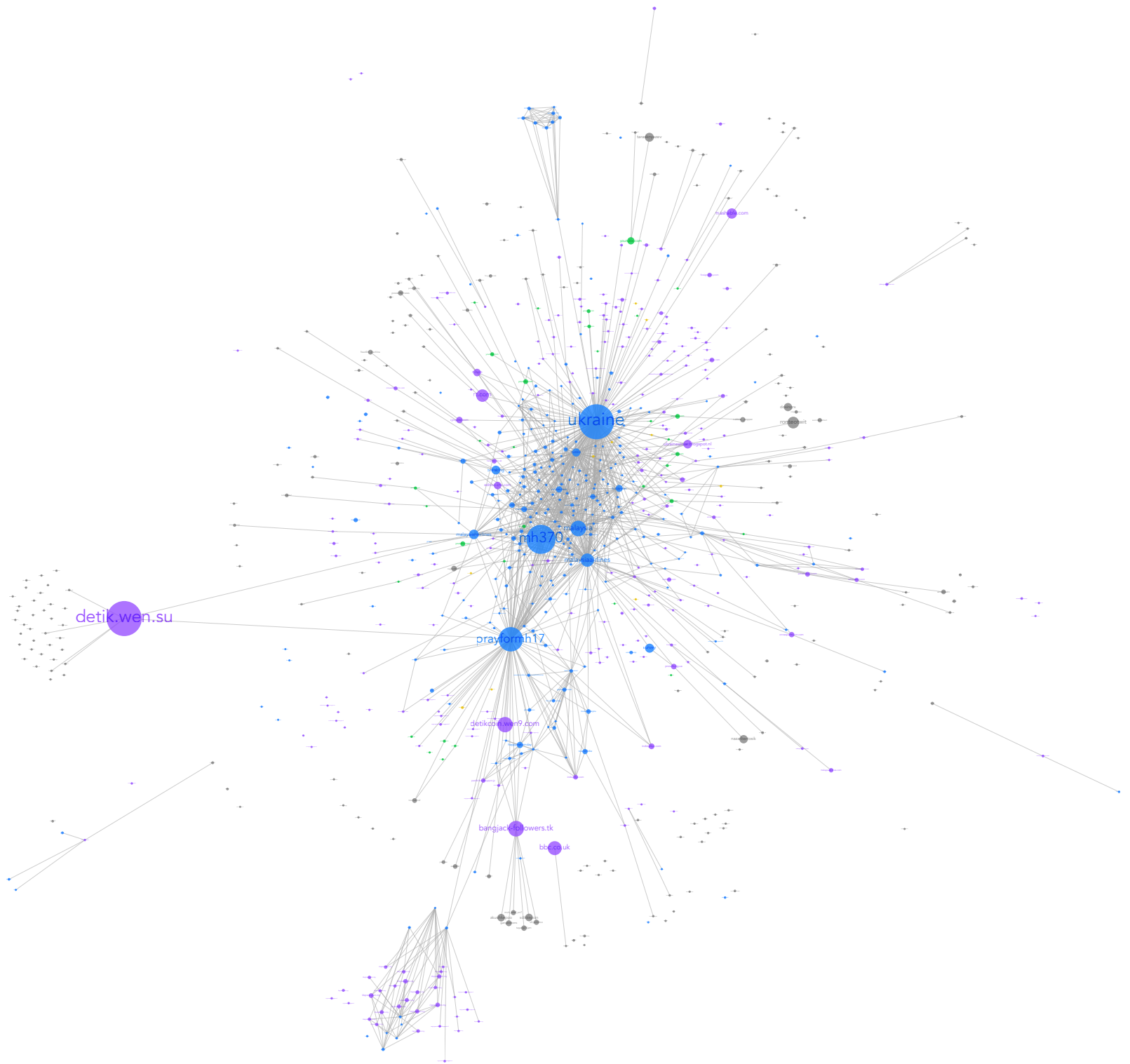
At this point all networks have been again imported in Gephi and appended one into each other. If duplicates were found (based on an equal value in the “label” column) they were merged maintaining all edges and summarizing the normalized frequency values (only one entity type per file had the value, therefore the sum was between one correct value and two empty cells).

The data table obtained served to visualize the bipartite network, on which the ‘Giant Component’ filter has been applied before running the algorithm’.

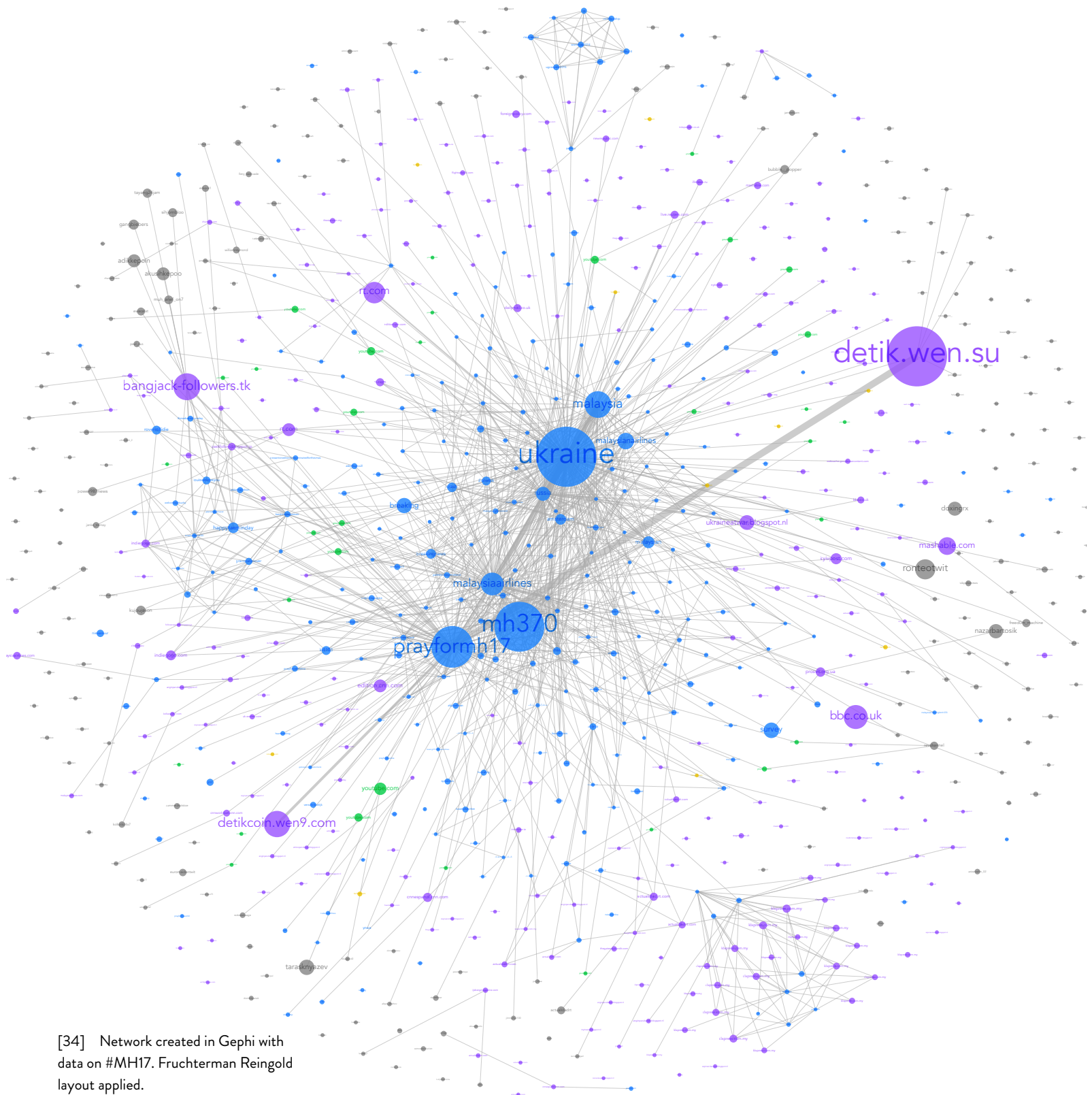
Since it is possible to reduce the data to specific timespans, it is possible to assign a different value to the entities in view to reflect the changes over time. Unfortunately this is not possible for values like retweets or favorites, which twitter releases in its final count but does not provide any track on how the value changed over time. Potentially however, it is possible, Twitter definitely has access to these data.

Further analysis can be done in order to reflect different kind of information. For the purpose of this work I left out these options to focus on one only, since the aim is to verify the usefulness of the visual approach as a method for exploration.

The last part of the visual mapping consists in the creation of the visual structure. In the network case the disposition of elements in the space depends on the spatial algorithm chosen. I’ve tried many different ones and based my choice mainly on the visual outcome. In particular, I have found these four algorithms (each



[33] Network created in Gephi with data on #MH17. Yifan Hu layout applied.



[34] Network created in Gephi with data on #MH17. Fruchterman Reingold layout applied.

available within the Gephi software) to produce interesting results: Force Atlas, Yifan Hu, Fruchterman Reingold, and OpenOrd. Mainly two components were used in the final choice: disposition of elements and overall required space. This is possible because the dataset is the same, so comparisons on the resulting representations are totally correct.

Force Atlas creates a various distribution of elements, where some groups are far from the central part, but overall it does not require too much space. The *Yifan Hu* is similar to the first, however it seems to push quite far even important nodes. The biggest value of the *Fruchterman Reingold* is clearly in its shape: this algorithm always disposes nodes to have a final circular shape. This is very good for the exploration, because it helps not to loose awareness on where one is when very much zoomed-in. On the other hand it looses visual separation of elements, making more difficult to identify patterns. Last the *OpenOrd*, which is recommended for big amounts of nodes. Even if it takes a lot of space, it works very well in the distribution and use of space, emphasizing distance between elements.

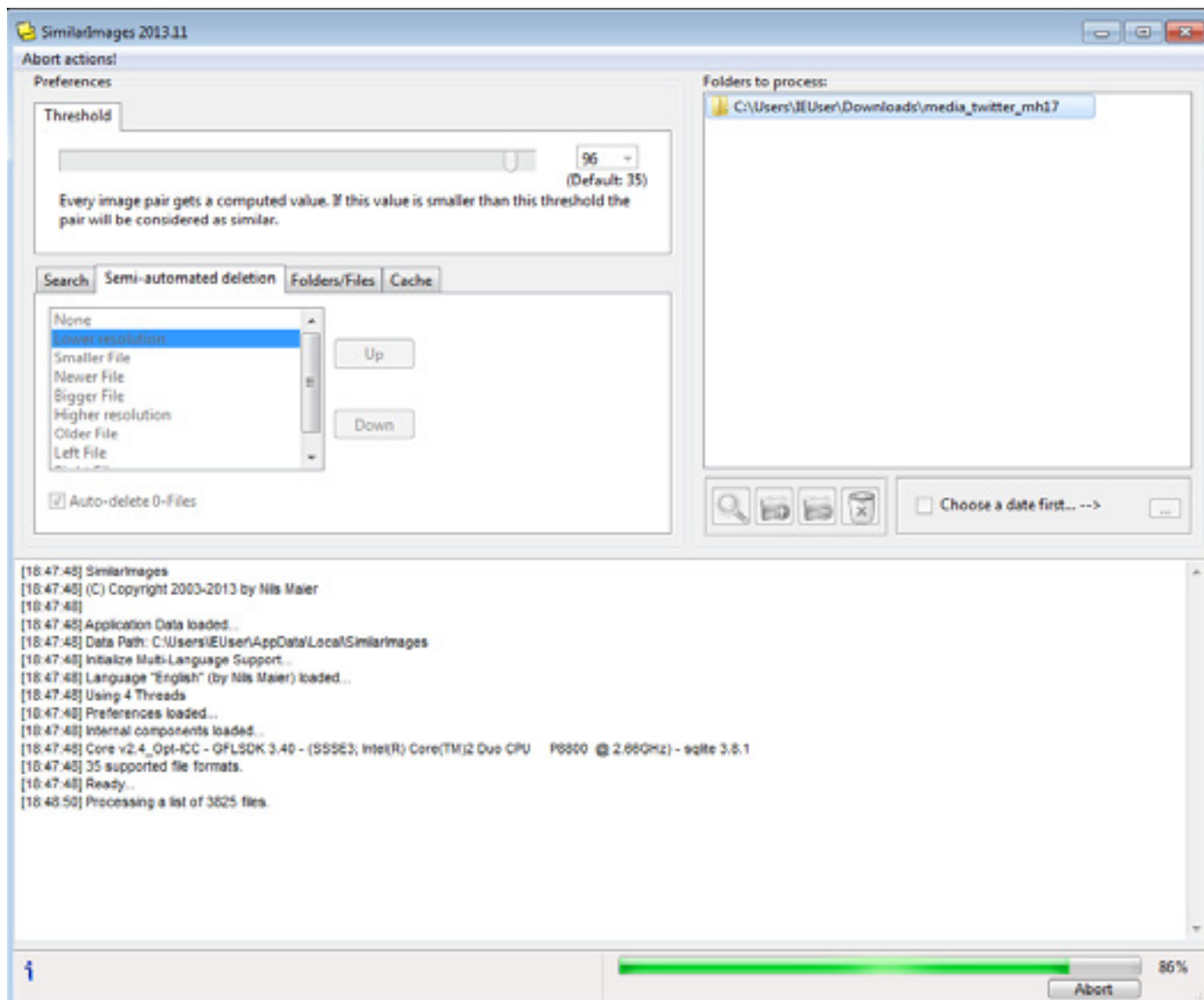
Concerning images an additional work on data has been done before the realization of any visual structure. The main issue with images is that users, whether they are citizens or professionals, often post the same content: an original image is taken, copied and sometimes modified (cropped or with addition of text or marks) and then reposted. This habit creates a lot of noise in the space by multiplying and splitting the conversation in several and minor ones. Browsing this media content poses two problems: firstly it creates an echo of duplicates, which require more time to go through; secondly it makes more difficult to keep track of images as belonging to a single entity. I thought of an innovative methodology to being able to treat even images as hashtags or URLs. To be more precise, I see the option to do not consider each image as a single entity, but to lead back to the original image (or at least to one image) all the variants users created of it.

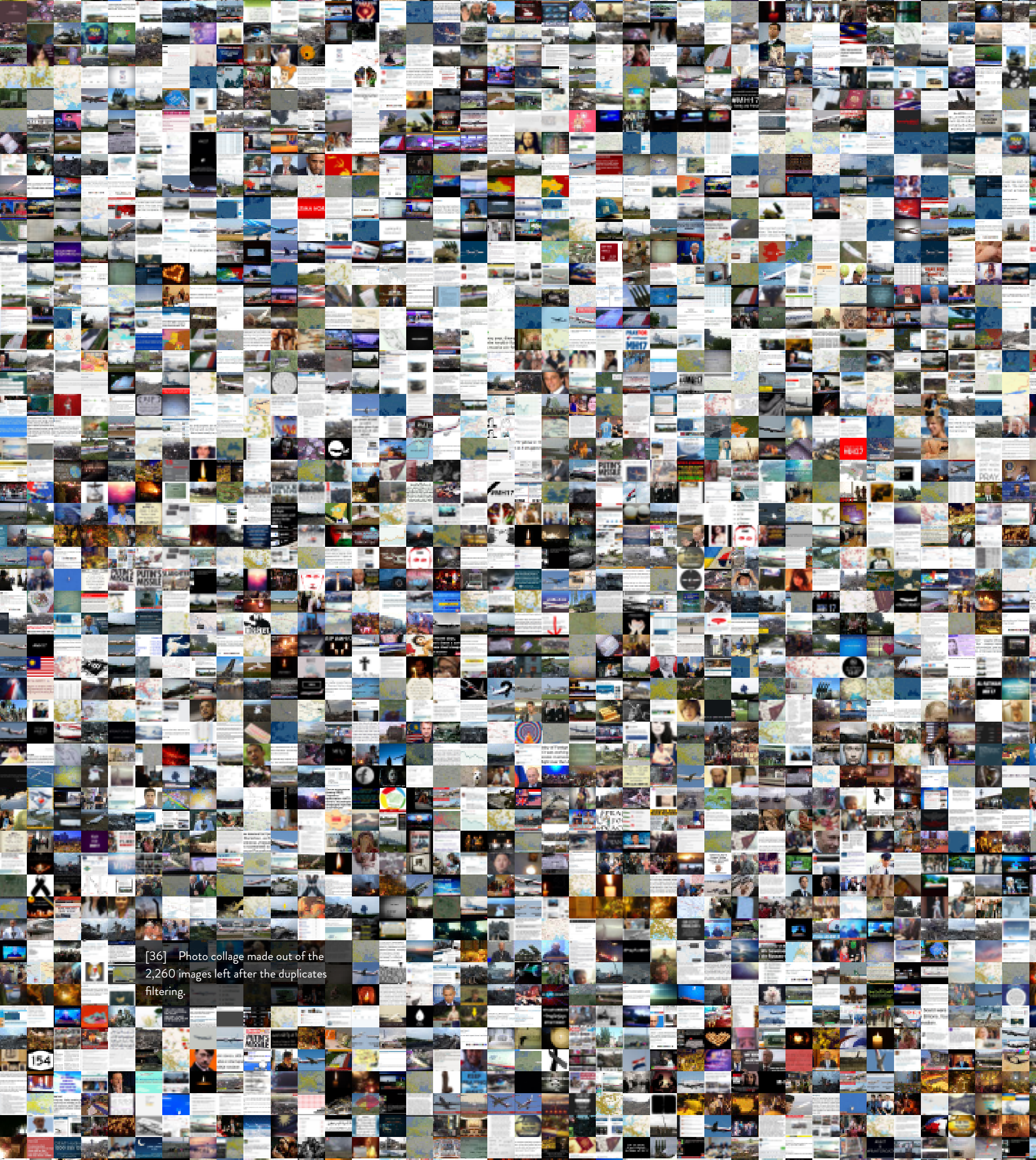
To do so there is need for an images analyzer algorithm able to bring together similar images. It would work as the same as *Google Search by Image* or as *TinEye*,

both image search engines. For the purpose of this work I looked into the market of duplicate pictures finders. I found *SimilarImages* to match all my requirements: freeware and with the capability of automatically delete found duplicates while keeping a log of what has been deleted and in favor of which.

From the TCAT tool I download the list of media URLs, so that each URL is linking to the actual image format file. Through the *Mozilla Firefox* open-source extension *DownThemAll!* I was able to automatically download the list of images on my

[35] Similar Images interface.





[36] Photo collage made out of the 2,260 images left after the duplicates filtering.

hard drive; I got a total of 3,918 files. I launched the images comparison analysis on SimilarImages setting a threshold of 96/100 (previous tests it demonstrated this value to match correctly images even when slightly modified). After finding matches, the software deletes all duplicate except one item, which will be the one with the higher resolution. With this method I was able to reduce the number of images to 2,260 (57%). SimilarImages creates a log with the names of the files that deletes and the one that stays in their place. A last work on data consisted in keeping only the list of the remaining files and counting their matches: this becomes the value of popularity of that media.

UGC have the possibility to embed information that before social media was not used: geo-location. Most social media platforms in facts give users the option to say where they were in the moment of sharing, and using the GPS technology this information can be very accurate. In my database geo-location for those tweets is shown as latitude and longitude values, easy to map on a world projection. In an early stage of the project I wanted to show how the topic is spoken around the world, giving this way a help to researchers on how to identify areas in which the topic is more trending. In doing so two parallel options came to my mind: geo-located tweets and location based on bio. While the first one is very precise (indicates a single point on a map) and secure, the second method needs to rely on what users say about themselves. Other ways to detect a user location might be to check the time offset, which can work as indication of the part of the world from which he/she uses the platform. Even though most of the times this is correctly chosen (with the selection of the capital of the country where the user comes from, in example Amsterdam and Rome both are in the same time, but they are selectable as different offsets), it happens that some users choose an the time from a different country in the same offset, making the information not reliable anymore. For this reason I decided to not go further with this option; however if a more reliable way to define the country of living, some interesting patterns can be detected and worth to represent. A good way to visualize these

data would be to use a color transparency or value to fill areas (such as countries) based on the percentage of active users in the conversation. To make the visualization more useful, an innovative idea was considered. When dealing with international topics, such as the MH17 case, the visualization would not distinguish between areas of generally higher activeness compare to the others. For example the area of New York, because of its population density, would probably be one of the hottest areas in an activity map, but this might be because of the normal activity of users in that area. While I would argue that by normalizing these values by number of users living in that area, would better show when an area is surprisingly more interested in the topic.

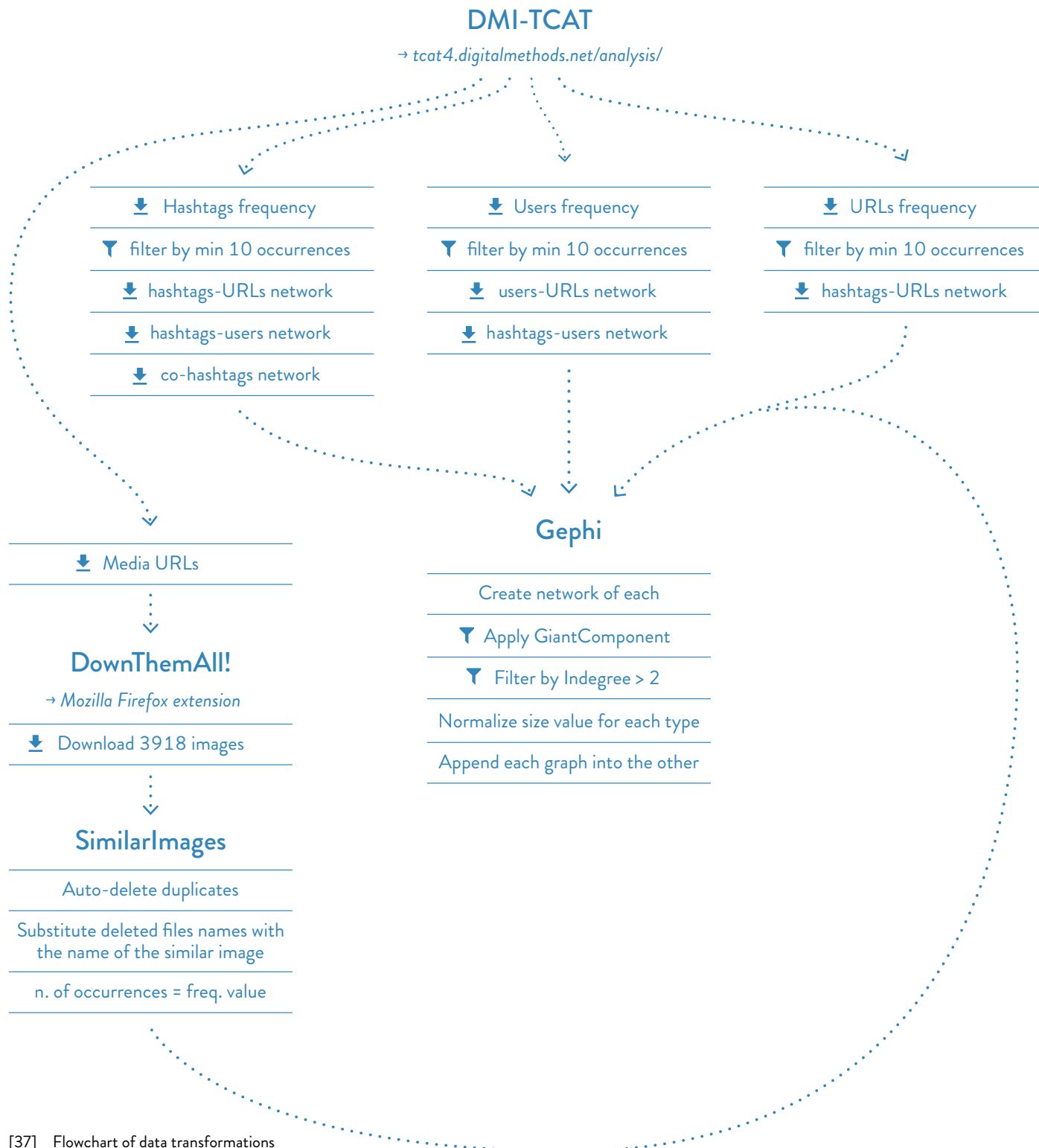
Due to not reliable data, geo-located tweets became therefore the only data shown in the Activity Map. Differently from the network visualization, in this case the minimal unit is the tweet rather than the entities. For this reason each point on the map visualizes a tweet, placed in the location in which it was shared.

To do this I exported from TCAT only tweets with embedded geolocation and imported the data in *TileMill*, cartography software.

Since the tweet is the point, the values that can be represented by visual variables are those related to the tweet, namely retweets, favorites, and time. It is than possible to filter the tweets to map based on entities, for example showing only geo-located tweets using a certain hashtag.

I would argue that this representation could provide mainly two kinds of insights: awareness of where a topic is hot regardless countries' borders, and help in the images and videos verification process. For example it is extremely useful to check the exact position from which an image was taken, and in this visualization the information is given within context.

The last representation wanted to focus on entities as seen out of their relations which each other. In particular I find interesting to see changes over time especially when the research spans over long periods. For this reason I started some work towards the identification of recurrences in sharing of the same con-



[37] Flowchart of data transformations process

tent to find how it changed over time. This is particularly useful to see if content became popular fast or slowly and when in time this change happened. To do so I used the export of each entity frequencies, to have an overview of which are the most shared content and how many times this occurs. I then used Microsoft Excel with the function 'Vlookup' to return the time in which each entity was shared. Before doing so, since the time value of the exported tweets specifies even the exact second, I grouped together times in timespans. For this project, with a database of tweets from a total of 10 hours, I picked 5 timespans (in other projects I imagine this selection to be automatic and relevant for the duration).

Finally I created modified the data table in order to be visualized in a streamgraph, which allows to detect streams of content over time, where each column is a timespan and entities are sorted by frequency. The streamgraph has been created in *Adobe Illustrator* with the *Scriptographer* script developed by the *DensityDesign-Lab* at Politecnico di Milano. The goals of this representation were achieved.

VISUAL STRUCTURES: NETWORK

The visual structure of biggest value is without doubt the network graph. It is a bipartite un-directed network, in which nodes represent entities and edges represent relations. The space acquires an important value in this type of visualization, where the distance between entities is determined by the algorithm as an indication of the relation among the two; thus if two entities are close it indicates a stronger link rather than those who are far away from each other.

Proximity is therefore an important parameter when visually exploring the representation, helping to identify clusters of entities, which evidently have a strong relation with each other (or they present edges with the same entity).

The chosen spatial algorithm is the Force Atlas, for its good combination of small space required and good separation and grouping of elements. However it is important to keep in mind that this algorithm might not work as good with different topics. Further work would definitely require some testing with different datasets in order to find the algorithm that works generally well.

There are five types of nodes in the network: Users, Hashtags, shared URLs, shared videos, and shared images. Each of them is visually represented by a different color, in order to help to distinguish entities' category even when they are very small and reading the label is difficult. The color distinction also helps to see if clusters of nodes are of the same category or not, and to see which type of content a user is sharing even before reading. The chosen colors do not refer to any specific meaning: their selection was only based on a palette with enough hue threshold in order to enhance perceptual differences.

Entities are also separated from a shape point of view: circles represent users, while squares represent content. This small graphic addition puts also on a perceptually similar level the hashtag entity with the attached content entities, as they can be all seen as component of the message (together with the lexical language). Users instead should appear differently because of their status of producers.

For the way data were put together, every node can have an edge with any node re-

ardless its category. Edges between nodes represent relations as co-occurrences of the two nodes within the same tweet.

From the network it is possible to gather many information, even from a quick look of few minutes. In this relies the best win that is gained from the representation. Considering the MH17 case, it is easy to see its main connected topics, which are the hashtags: ‘#Ukraine’ is the biggest node, followed by ‘#MH370’, ‘#Prayformh17’, ‘#Malaysia’, and ‘#MalaysiaAirlines’. A bit smaller, but at the center of the network, we can see ‘#Russia’ and ‘#Donetsk’, other entities that refer to the location of the crash. The most interesting part of the network occupies the top part of it, with all the hashtags referring to Middle East’s countries: ‘#Gaza’, ‘#Israel’, ‘#Afghanistan’, ‘#Syria’, ‘#Bordercrisis’, ‘#PalestineUnderAttack’, ‘#saveGaza’, etc, ... By viewing few tweets with these hashtags it is suddenly clear what the discussion is about: the crash of the plane is moving the topic of discussion away from the war in MiddleEast and some users are asking to do not stop talking about it.

Looking at the purple nodes we can have an overview of the most shared URLs and their position tells to which topic they are most related to: ‘kyivpost.com’ is half-way between #Ukraine and #Russia and close to it the ‘ukraineatwar.blogspot.nl’, which is linked with ‘#Amsterdam’ but closer to ‘#Russia’ (probably because of its content). Again, there is part of the network talking about Americans, citing Obama and the US. Relatively close to it other politicians are nominated: #Poroshenko and #Putin (interesting are also the hashtags #stopPutin and #stopRussia).

Another help from the network is given when clusters of entities are far away from the discussion: it helps to identify why and which topics are together. It is the case for example of the cluster of hashtags at the bottom. At the center ‘#worldcup’ is pretty explicative, and the only connection to the rest is with ‘#Ukraine’.

The bipartite network visualization works very good in accomplishing its goal of giving a macroscopic view of the conversation. Just a glance at the vi-

sualization can easily tell which are the most used hashtags and how they relate each other, which are the contents shared together with their use and where users put themselves in the conversation (based on what they talk about and what they share). Therefore it is much faster to grasp information in this way and I would argue that it wouldn't be possible to get the same grasp even by browsing through all the generated tweets. The combination of computation, filtering, and representation can instead serve as a support in the exploration and investigation in a more efficient way. As for the gathering of the data I eventually had to scroll the whole list of 53,356 tweets, by doing this I was able to accomplish two things: put myself in the role of a journalist who browses through the list of tweets, and starting to build some ideas about how the discussion took form in the first hours after the crash. I have noticed that this method has enormous flaws that I will present here. It presents the information in a reversed chronological order, requiring the user to make a reverse thinking in which everything new he will read or see (scrolling down) actually happened before. Reading single posts gives a more specific idea on what single users are saying. This however might be very personal and quite irrelevant if the context is not clear. The order in which posts can be browsed poses again another problem: it loses any kind of grouping, requiring the user to mentally make associations between a large amount of posts of which he/she might already have lost track of. Understanding which are the most related hashtags, for example, is not easy if no quantitative analysis is taken into account. Again, having such a huge amount of information (which requires more hours to browse through than the time it took to be generated) makes it more difficult to detect changes over time.

However putting together so many categories in one network can produce visual noise in the visualization, making more difficult to see relations between two specific categories or among entities of the same category. For this reason I decided to have multiple network visualizations, each showing a different combination of categories. For example the users network visualization will not show

co-occurrence, but mentions. In this way the graph will display relationships in a more concrete way, helping to identify communities based on how they reach each other rather than just about the content they talk about.

All these secondary visualizations serve to focus the attention on relations between entities of choice. The data table created allows to choose which and how many entities categories to keep in sight, spanning from five categories to one only.

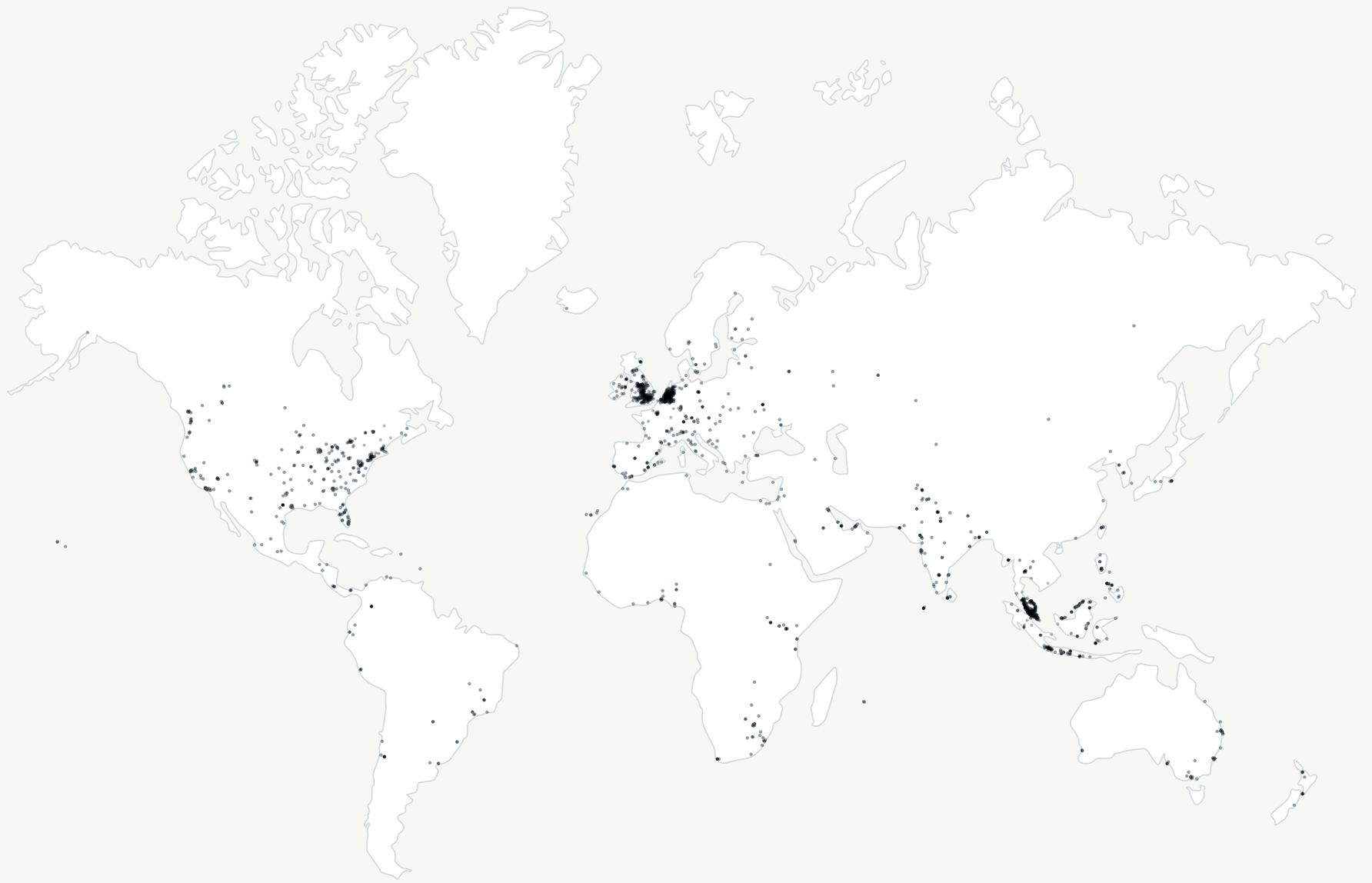
VISUAL STRUCTURES: ACTIVITY MAP

The activity map visualization serves in the exploration of geo-coded content. This specific type of representation do not present particular innovative component: many tools online offer a view to map on a world projection all tweets resulting from a query. However I strongly believe that this part is very important for the completeness of the work and an essential part in a UGC exploration.

The activity map pins a dot in the exact location from which a user tweeted. The size of the circular dot is instead related to variables that could be chosen like number of retweets or favorites. I find important for this feature to be implemented in the tool since part of the goal of the project is also in providing an interface for the exploratory process that do not require the use of multiple platforms to be accomplished. In addition, its implementation provides a different way to explore, based on a geographical matter. While working on a very limited amount of UGC (for the MH17 case only the 5,26% was geo-coded), the map shows precise locations or an overview on where in the world a certain discussion most takes form.

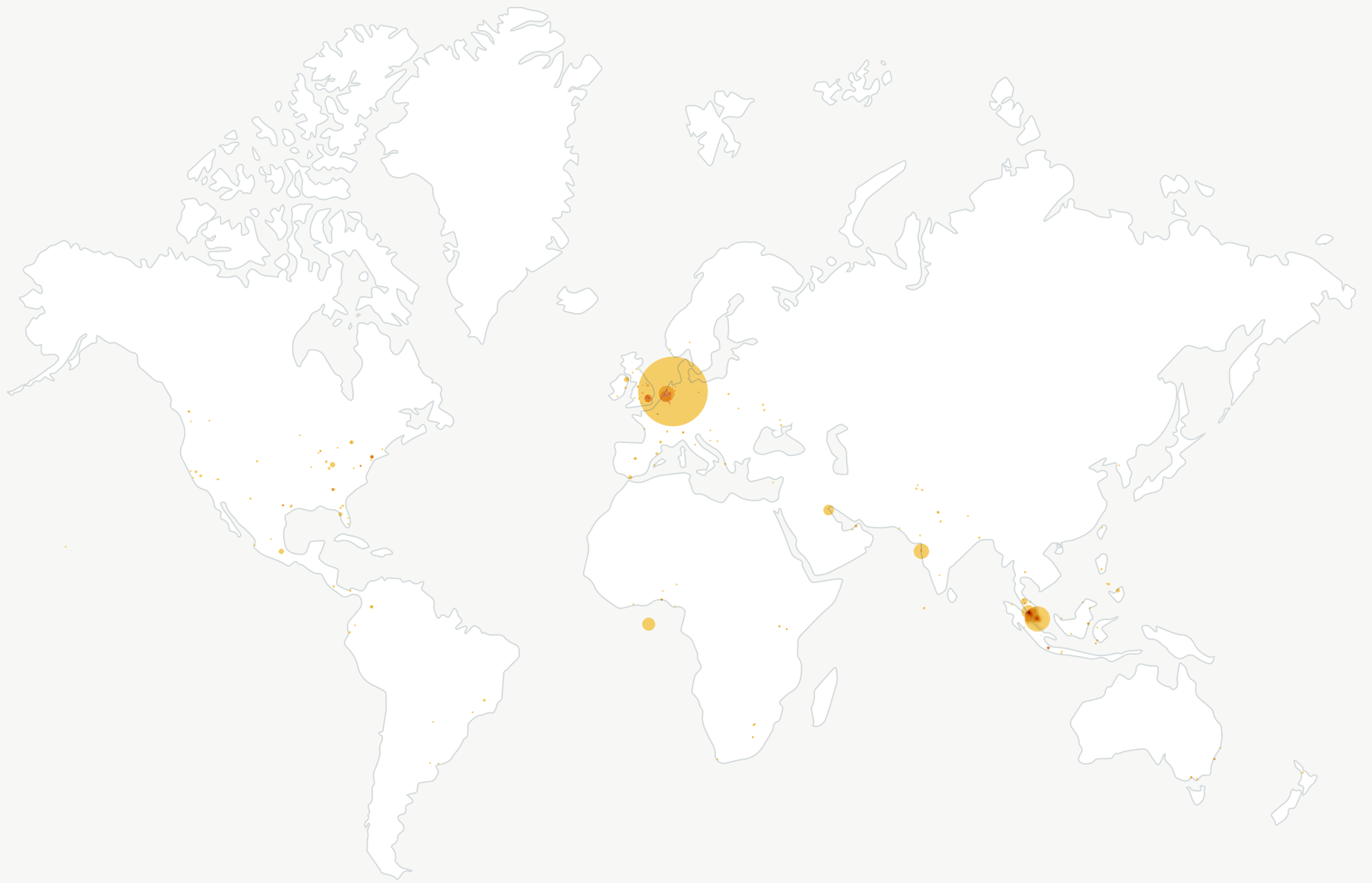
Since each single datapoint represents here a tweet (the whole message) and not those single parts that forms it (entities), the main visualization maps activity by showing most active parts of the world. Of course the assumption by which empty parts of the globe are not taking parts in the conversation is totally wrong. In facts it might happen that in some locations users are less interested in geo-coding the content they publish, meaning that the already low percentage of geo-coded content is with high probability not equally distributed.

However I think this representation presents some innovative features, such as the possibility to see changes over time or the possibility to view specific kind of content only. A map to see only tweets that embedded a picture is for example highly useful, because it might be a first help in the verification process. In specific cases, can also provide a fast way to see in which specific locations an event has been mostly recorded, highlighting new ways for further investigation or to confirm previous hypothesis.



From the MH17 case, we can gain some interesting information. All tweets let us see densities. We can clearly identify three highly populated areas: United Kingdom, Netherlands and Malaysia (USA is generally populated but evenly spread). Those can be quite obvious, since the latter two are both highly involved in the case (the flight was Amsterdam-Kuala Lumpur). Interesting are two unexpected things. The interest of UK seems to be due to its closeness to the Netherlands (but does not seem a sufficient condition), further investigation reveal that 10 people on board were British. The second unexpected thing is the almost emptiness of the area around the crash site, which is located at the northern part of the Black Sea. I can imagine this area to populate exponentially when tweets of the upcoming week were integrated.

Another use can be done by mapping only geo-tweets with a certain hashtag. I tried with '#Gaza', expecting tweets in the MiddleEast. However those tweets are very few (they are, after all, related to a side topic in comparison to the crashed plane) and mainly located in Europe. This can fastly confirm or defeat some hypothesis.



VISUAL STRUCTURES: CONTENT TIMELINE

The content timeline represents the third and last visualization of the tool. Its goal is to provide a summary of the most relevant entities in the database and how they change over time.

To do this the alluvional diagram works at best by showing a list of entities sorted by relevance. It presents multiple lists, each for a different time-span of the research in which the same entities are re-sorted to reflect the situation of that span of time. In this way the graph can be read both vertically (the order provides information about relevance) and horizontally (discovering the most relevant entity over time). The visual structure of the alluvional diagram adds a visual clue by representing the relevance value as the height of a shape that flows through columns always aligning with the same entity. The graph shows indeed a separate flow for each entity and by following it is possible to grasp if that entity is climbing the rankings, staying at the same position or going down.

I set the value that defines the height of the flow as the number of tweets with that entity. That can be changed with any other numerical data of interest, which will consequently change the information that can be grasp from the visualization, maintaining the reading process the same.

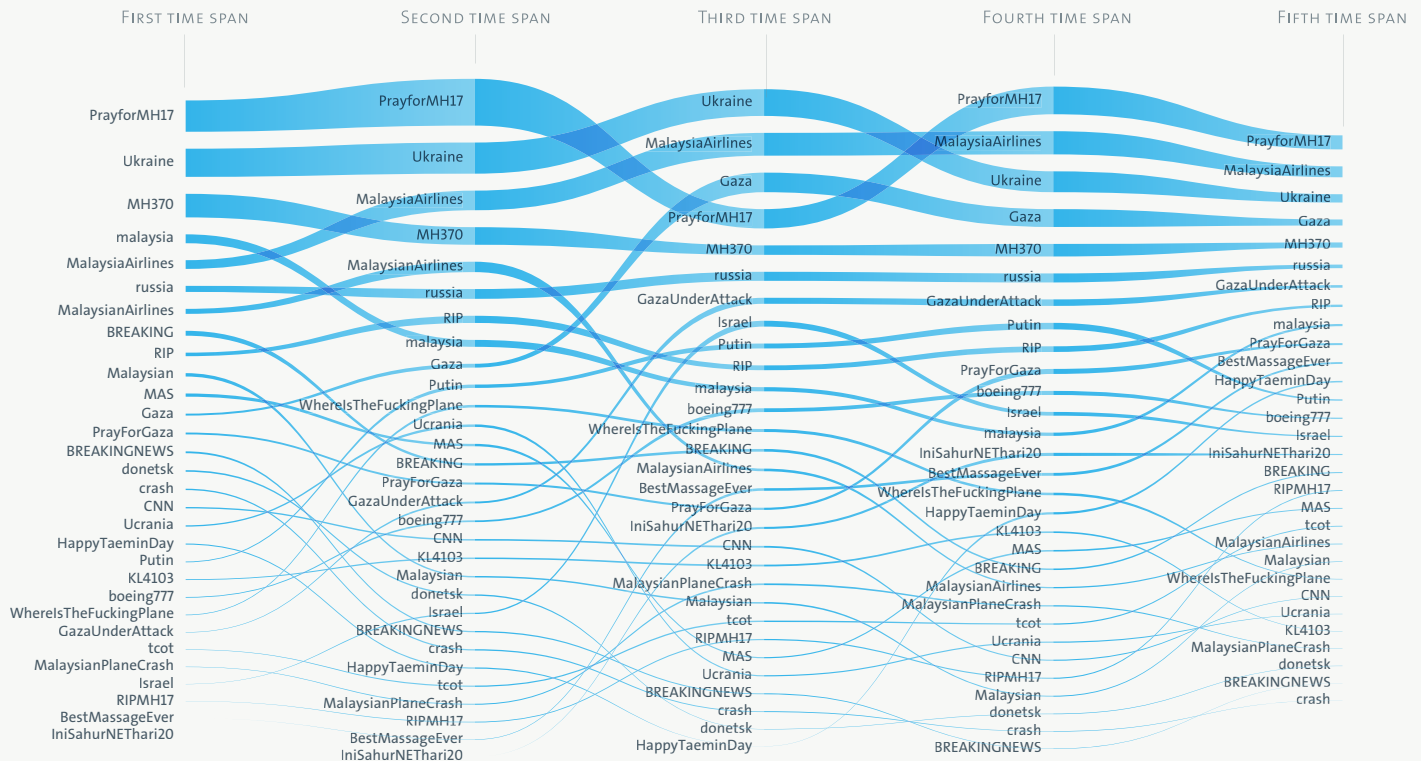
This visualization provides interesting insights about the MH17 case. Starting with users it is noticeable that there is one user that heavily holds the discussion: @WitFP. In just 10 hours of data, the profile tweeted 215 times with the #MH17 hashtag, an extremely high amount. This might be an indication of spam, but opening the Twitter page reveals to be a user created appositely to talk about the case. Further exploration of its tweets makes it easy to understand the anagram of the username: '#WhereIsTheFuckingPlane' is the most used hashtag. Another interesting pattern is that most users that are at the top of the ranking in the first timespan, drastically drops in the second to leave the place to new users, which become more active in the discussion.

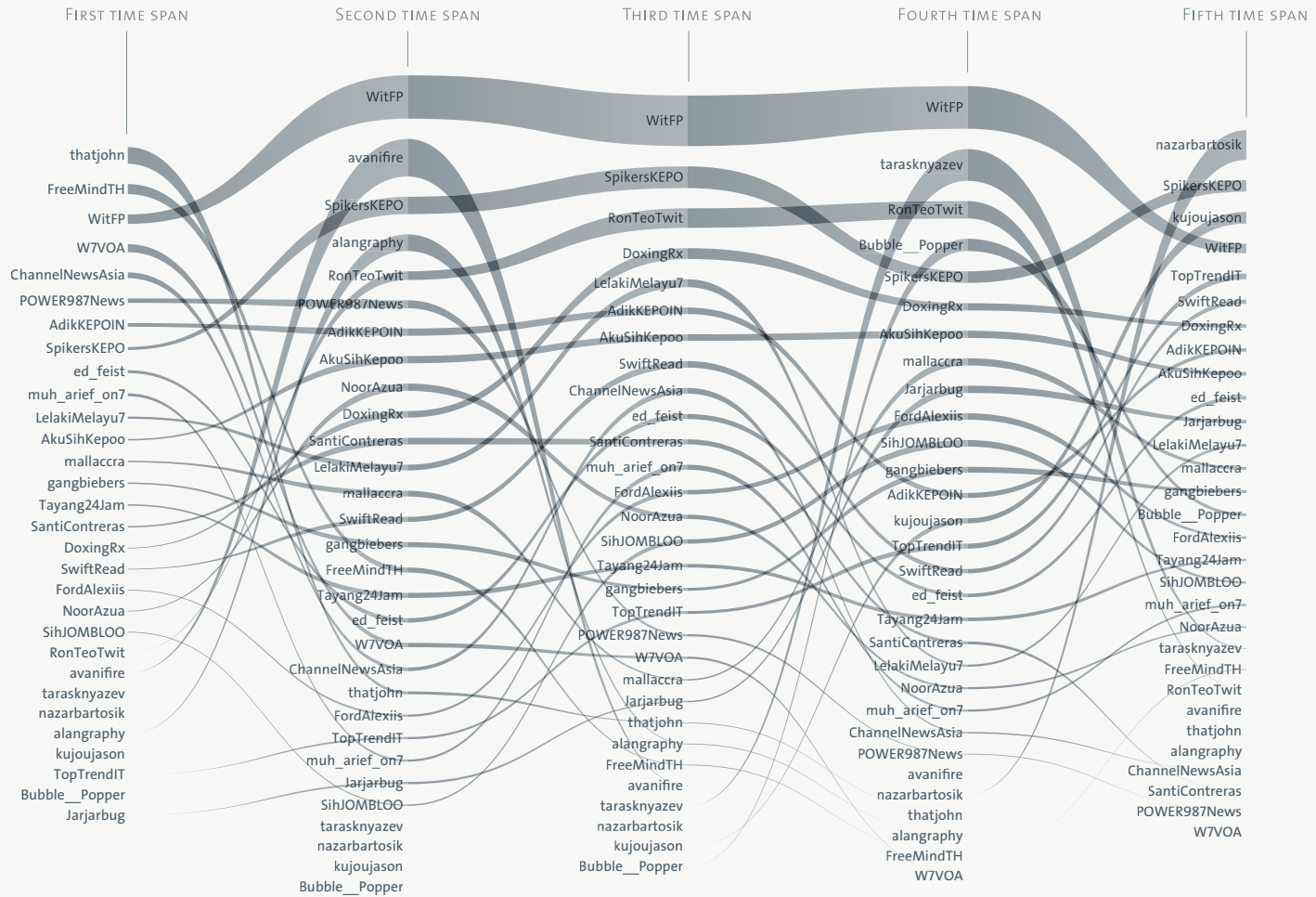
In the URLs timeline we find the ukraineatwar.blogspot.nl to be interesting for its 0 value in the first hours and to achieve a second position in timespan 3. It turns out to be a continuously updated and detailed article about the crash and its related information.

Concerning images I find this visualization interesting to portrait how the evolution of knowledge (or of perception) of a topic changed over time. In the MH17 case the images mostly shared within the first timespan are about smoke seen from a distance. In the consequently hours images from closer arrives and at last we have pictures quoting Najib Razak, Prime Minister of Malaysia.

I am certain this kind of visualization can portrait better more interesting changes when the project will cover a longer period.

Another useful section is the word used timeline. Together with hashtags, this is the only visualization to work on a semantic analysis. In the case of MH17 the selected timespan is too short to actually see interesting changes over time. However a small change is still in sight: in the first two columns the discussion is about the event itself, while in the last is already more about families and prayers.



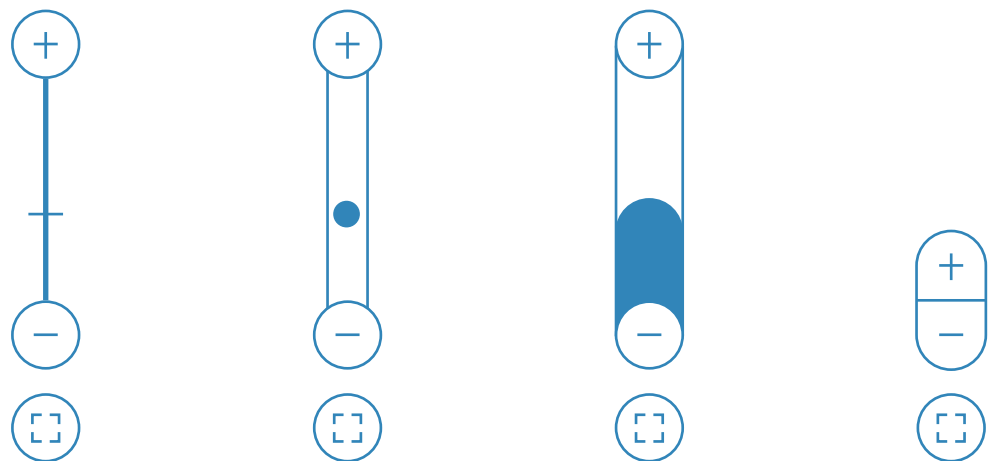


VIEW TRANSFORMATIONS: INTERACTION

Designing Information Visualizations by definition asks to design interaction possibilities. In particular, aiming to be an exploratory and investigative tool, the interactive component of the project becomes pretty significant as amplifier of cognition and in order to give the freedom that every exploration deserves.

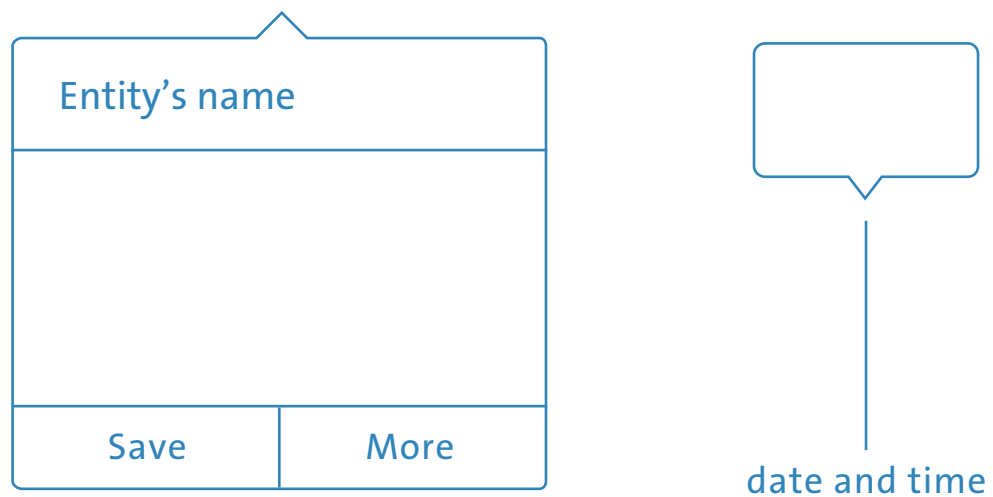
In this paragraph I will highlight the options designed for interact with the visualization that will be part of the UX and UI design. To focus more on the value of the options designed and less on their visual appearance, the images will present components in wireframe style. Their final appearance can be seen in the last chapter of this thesis.

I would start with the more subtle yet essential zoom interaction. Both in the Network and in the Activity Map visualizations the possibility to interact on the scale is not only useful, but it is needed. The two visualizations work similarly in the way data is represented (they can both be seen as maps) and by increasing the scale more details are shown. Since all information cannot be shown at the farthest zoom possible (which provides the bigger picture) the zooming-in interaction works not only as an abstract technique, but in a small part works as a details on demand approach too. In the network visualization for example, where the label



of each node is scaled together with the size of the entity, the name of nodes is displayed only when it reaches a sufficient font-size to be read (8 px). By zooming-in the user increases the scale, making even the smallest nodes legible. From the moment that is up to the user to decide where to zoom, details are shown for the entities he is already interested.

Within the details-on-demand technique, the main interaction option takes the form of the tooltip on hover. The tooltip is an additional item that is shown only when a specific item is hovered (a cursor is required, the same interaction would require a tap on a touch interface). The tooltip will be shown on top of everything else, temporarily hiding anything behind. Most of the times a tooltip features a small pointy shape that highlights the hovered element in order to emphasize the relation between element and additional information displayed. In the tool this interaction is strongly used as a main way to have insights from any object even before deciding to investigate further (as a matter of fact the hovering interaction is the one that defines whether and item needs more investigation or not). In the network, as well as in the content timeline, the tooltip shows information about each selected entity showing for example the complete name, a preview of the article or the profile picture of a user. Hovering on each dot in the activity map will





Hashtag

Hashtag

n. of times

n. of distinct users

Show on network 

Most related hashtags

Most active users

Random example

Username
 Text of the tweet.



Activity Map

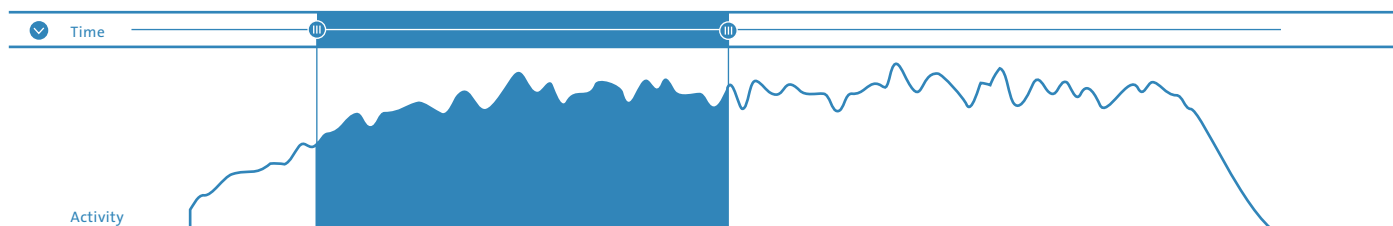
instead show the entire tweet. The same technique is used hovering on any part of the timeline, providing precise data about the selected moment.

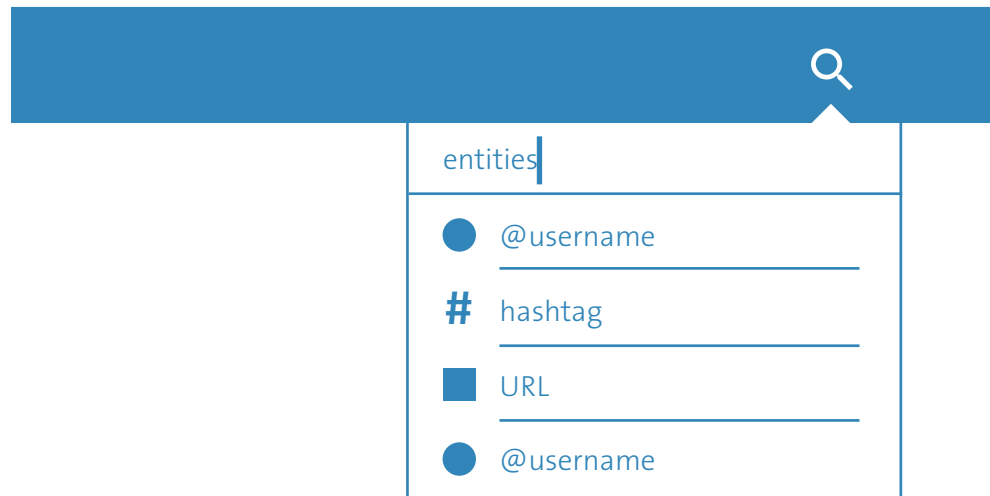
Talking about details-on-demand every tooltip features a ‘more’ button. This brings the user even more details about an item of interest and is brought in sight only when requested. I made use of the concept of cards. To each entity is assigned a card with all details such as how many times it was used, when the first time, which are the most related entities and how it is its use on a map. The card is shown as an overlay that blocks any interaction behind it. If the card has a long content to show it will feature an internal scrollbar, maintaining the card in position on the screen. In the header the most relevant information are shown.

To extend the interaction possibilities, every entity shown in a card (such as related entities) will be selectable and they will open up a new card, making possible navigation through cards.

Cards can be seen as the answer to the desire of further exploration, where the researcher can verify some information. If the information are not enough, or if it is necessary to further explore outside of the tool, a link to the actual entity on the web is provided so for example it is possible to open the profile page of a user on Twitter (this will open a new browser window).

The timeline is an essential component of the tool. First of all it fulfills several tasks: indication, representation, and filter. It serves as a reminder of how much of the entire timespan of the project is currently in sight by visually highlighting the portion in view. It is a representation itself, because it shows activity or share





frequency over time. It is also a filter, because the edges of the portion in view are scrubbers that can be moved over the timeline to extend or reduce the data shown in the visualization. Cards also implement a timeline. More precisely they modify the existing timeline to display the activity of the entity in the card rather than the activity of the whole project. The tooltip information is still available, while selection of time is not possible, since the card summarizes the entity in its completeness.

All three visualizations embed a set of tools: search, information, edit and settings. The *search* will let a user look for a specific entity (by name) within the visualization. If one is selected from the results, the visualization will automatically adjust to have that entity in sight in its selected state. This might help to check if a known entity (such as a user) is in the visualization and what might be its surroundings. The search works also as a location search engines in the activity map. This can help the researcher to see which content was shared in the surroundings of a place of interest. In the case of the MH17, for example, it might help to see if there was geo-located content in the area close to the crash (which together with the time selection can restrict the shown data, becoming a valuable method to verify). *Information* works as a legend. It provides a short explanation of what is in sight, how to read it and which information can be gathered. It is not really and interac-

tion techniques, but it is an essential component of the interface.

Edit is a fast access to filtering and editing options for the visualizations. For example in the network it is possible to choose which entities to keep in sight, letting the bipartite network to become a user-only one that will modify the structure of the net. Finally *settings* collects all advanced options in one window. From here it will be possible to select the filter threshold of the entities in view, letting move from the recommended view (which filters entities to ensure a not cluttered view) to a more or less filtered view. I imagined in this place all the other settings for the project,

Settings ✕

- Network
 - Activity Map
 - Entities Timeline
- Project
- Profile

Entities in view

- Hashtags Keywords associated with your research topic. Note that for best results the hashtag that identifies your topic research has been ignored. [MORE INFO](#) →
- URLs URLs linked together when talking about your topic.
- Images Images posted and URLs linking to Instagram and Flickr or to any image format.
- Videos URLs linking to YouTube or Vimeo or videos posted directly on the platform (e.g. Facebook or Twitter).
- Users Profiles talking about your topic within the time span selected.

Quantities

Show nodes with a value greater than: ▼

Explanation of how the value is calculated. For example etc...
Not that a lower value means more nodes in the visualization, making it more difficult to browse and cluttered.



Dossier

Saved items



Notes

 New note



Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.



Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna.

● *Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.*

such as the total timespan of the project or the user profile settings.


Overall other considerations have been taken into account during the interaction design of the tool. Mainly the need for a big canvas during exploration drew the line for keeping elements in sight only when needed. Hence all elements that take a consistent space of the screen have the option to be minimized, but always accessible.

The timeline, for example, can be minimized to a few pixels height, keeping however its property of being indicator of the selected timespan.


Another important part of the interface is the one given by the Dossier sidebar. It works as working space for the user: in this place he will find all the items he has saved (through the heart button available while hovering any entity) and a storyline where to place entities and notes. The sidebar is composed of two sections: the list of saved items; last the storyline with notes. The latter works as a vertical line that invites users to drag&drop the saved items from the list above, letting them to give an order that suits a storytelling rather than a chronological sequence. The sidebar is intended to be an extension of available interactivity options and in future developments could be definitely expanded to a dedicated area where all the collected information gets meaning together. I would argue that within this sidebar it occurs the transformation of information into knowledge, where the understanding of something previously unknown takes a tangible form, ready to be used or to be shared.

Final Design

My Research Projects ▾



Start researching
a new project



MH17
17 Jul 2014 – 18 Jul 2014

Home

Thumbnails of each project are aesthetic-oriented data visualization based on geo-location of content and type of content shared

Project

Network

Provides the overview:

- main entities in the discussion
- relations between entities

Activity Timeline

- Shows activity over time
- Work as a filter to adjust the visualization to a specific period of interest
- It's possible to bookmark events

Tooltip

- present on hover, it provides fast detail-on-demand

Activity Map

Shows geo-located posts:

- where is the topic trending?
- helps to look for specific entities based on the location
- save locations to see proximities

Card

- it's the place where the information reaches a granular level
- each card shows information relevant for the selected entity



Page



Component

Content Timeline

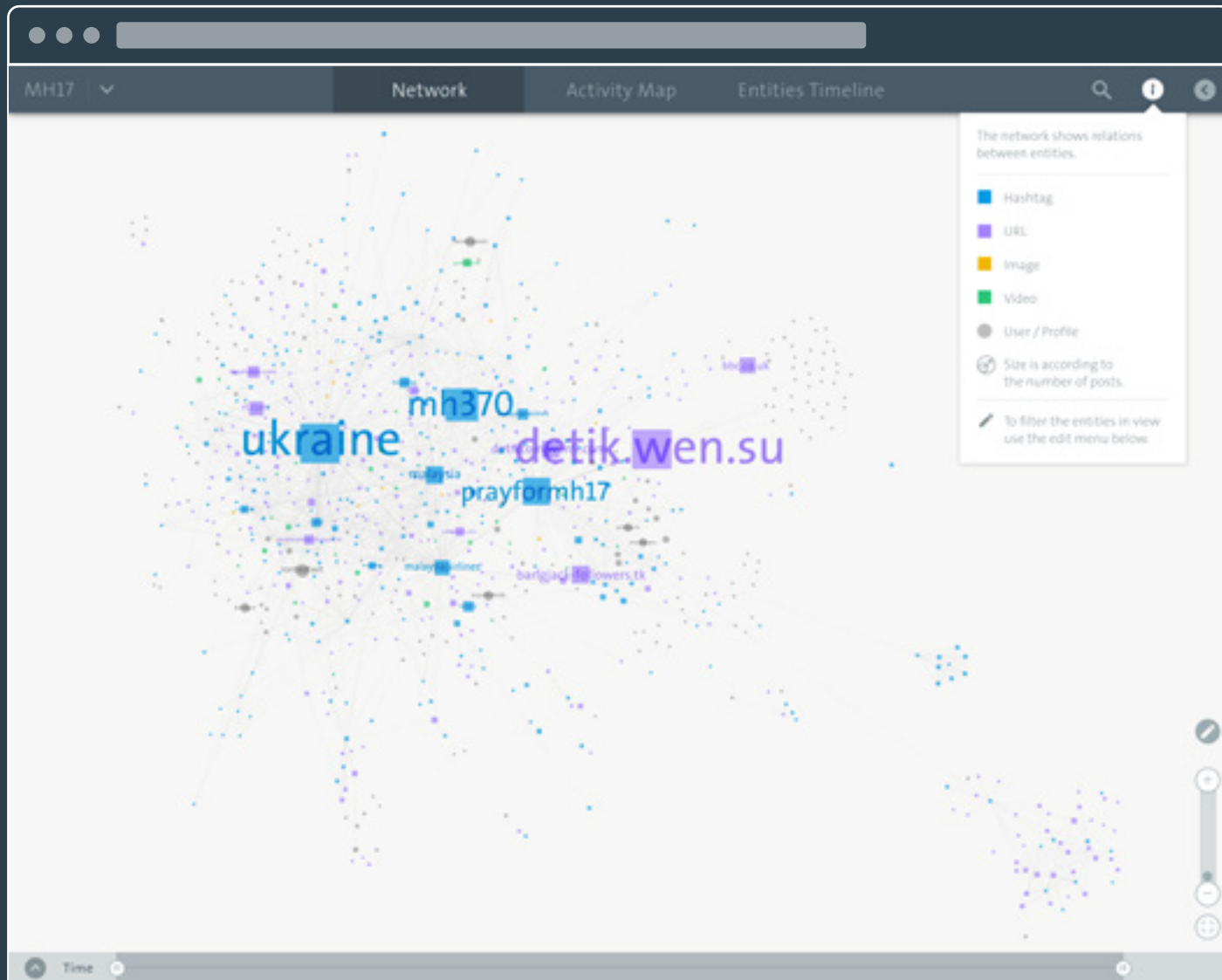
- Shows size changes over time:
- when did an entity started to be an important node?
 - see rapid changes in one view

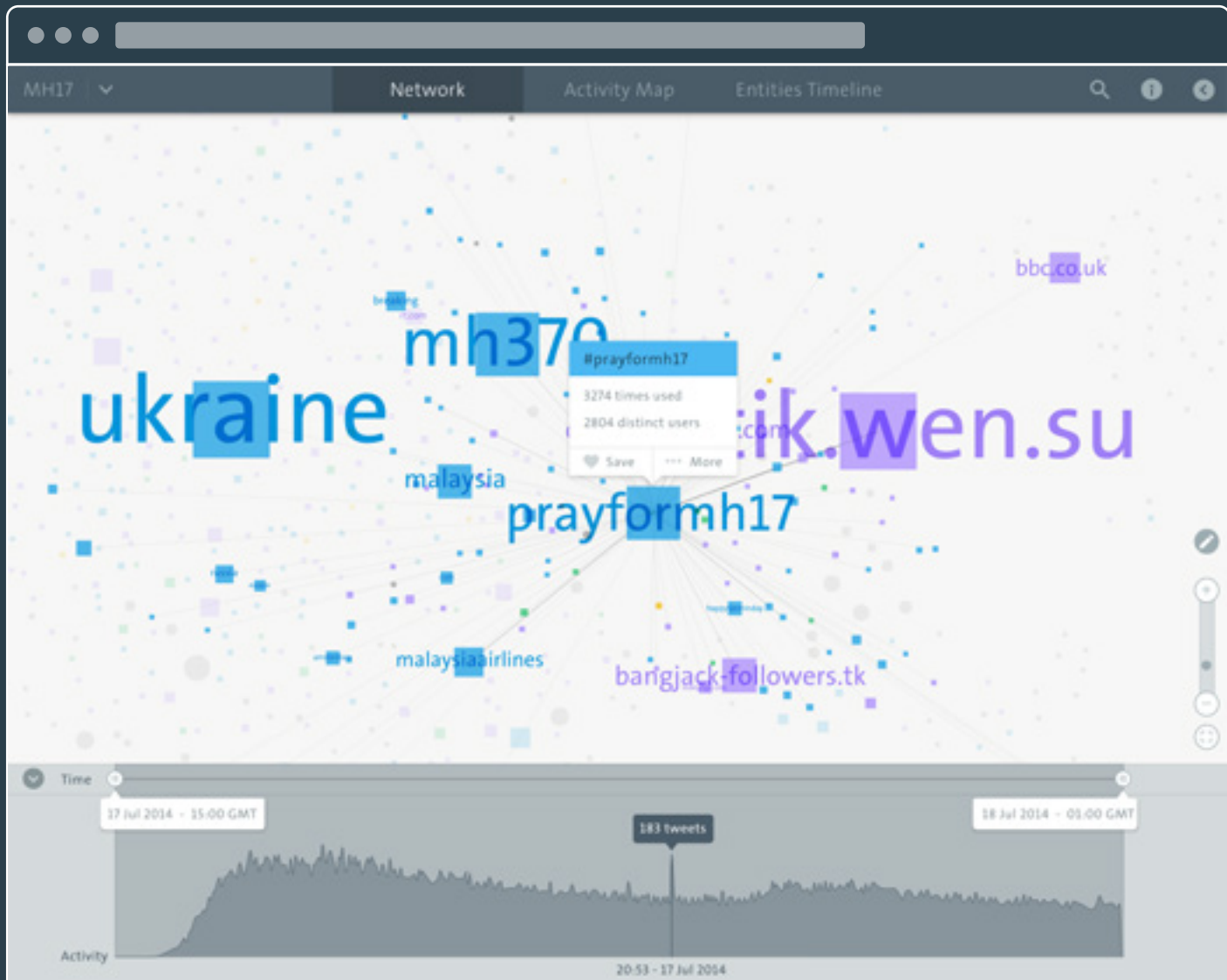
Dossier sidebar

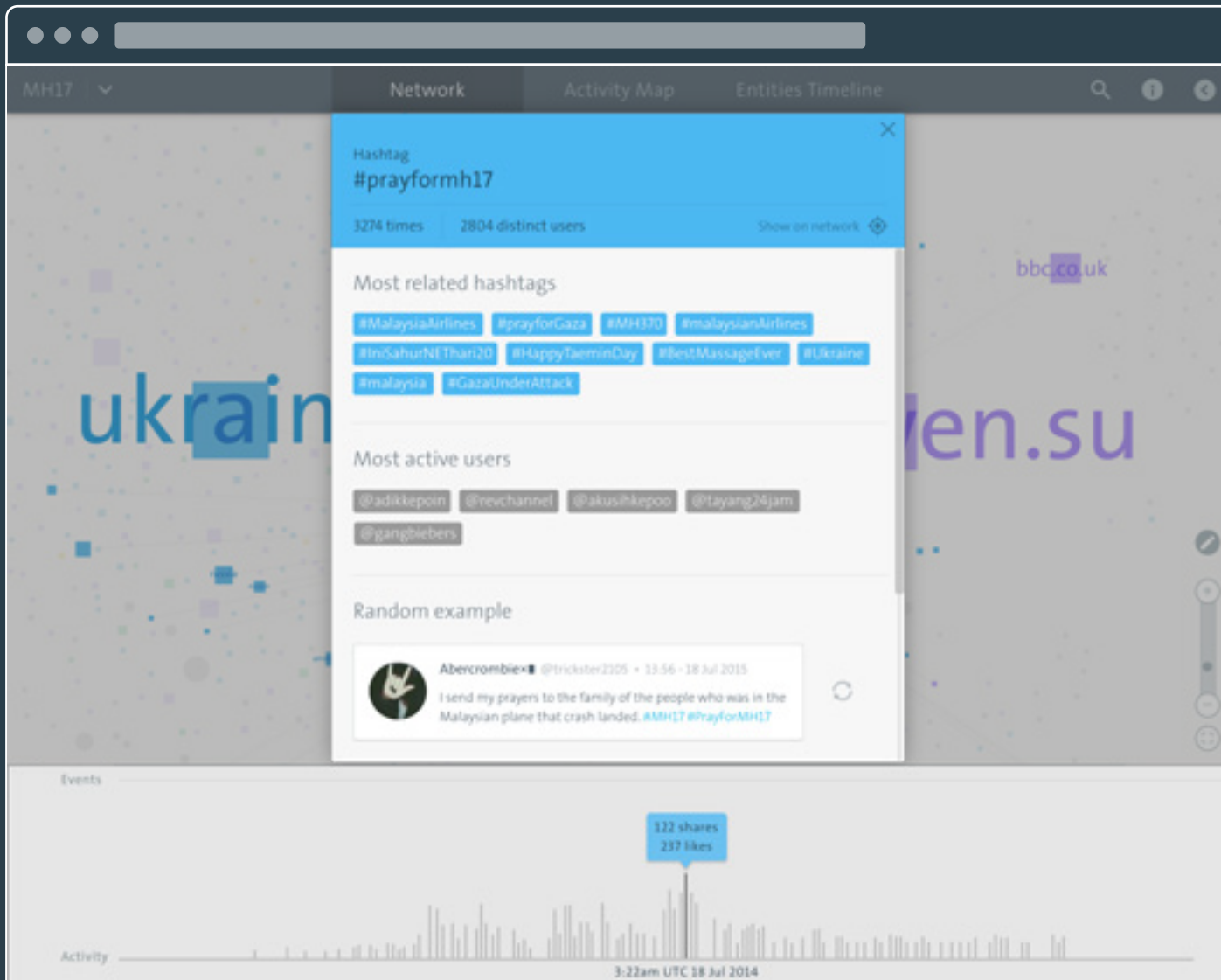
- provides fast access to saved entities
- includes a story-timeline on which to add notes and interesting entities

Settings

- includes all advanced settings to edit and customize the values that define the visualization







@marshacoll
Marsha Campbell
 246 followers | 621 following | London, UK

used #MH17
11 times

Used hashtags

#RIP #tragedy

Mentioned users


@trickster2105 @akusihkepoo @wireduk

Shared content

Why Planes Still Flew Over Ukraine Until MH17 Was Sh...

MH17 crash: Aids researchers heading to M...

Tragedy on another Malaysian Airlines plane...



at 00:30 GMT
18 Jul 2014
 Show on network

Ukraine

8 people aboard was... is obviously a dangerous... n fighting there for months... een shot down i...

Open link


at 17:31 GMT
17 Jul 2014
 Show on network

at 17:31 GMT
17 Jul 2014
 Show on network


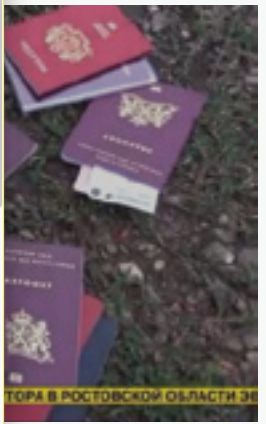
#MalaysiaAirlines #Ukraine #crash #shot

Most related users


@wired @revchannel @akusihkepoo



Visually similar images:

Visually similar images



17-07-2014 Instagram | 17-07-2014 Facebook | 17-07-2014 theguardian. | 17-07-2014 Instagram | 18-07-2014 Instagram | 18-07-2014 guardian

MH17 ▾ Network Activity Map Entities Timeline

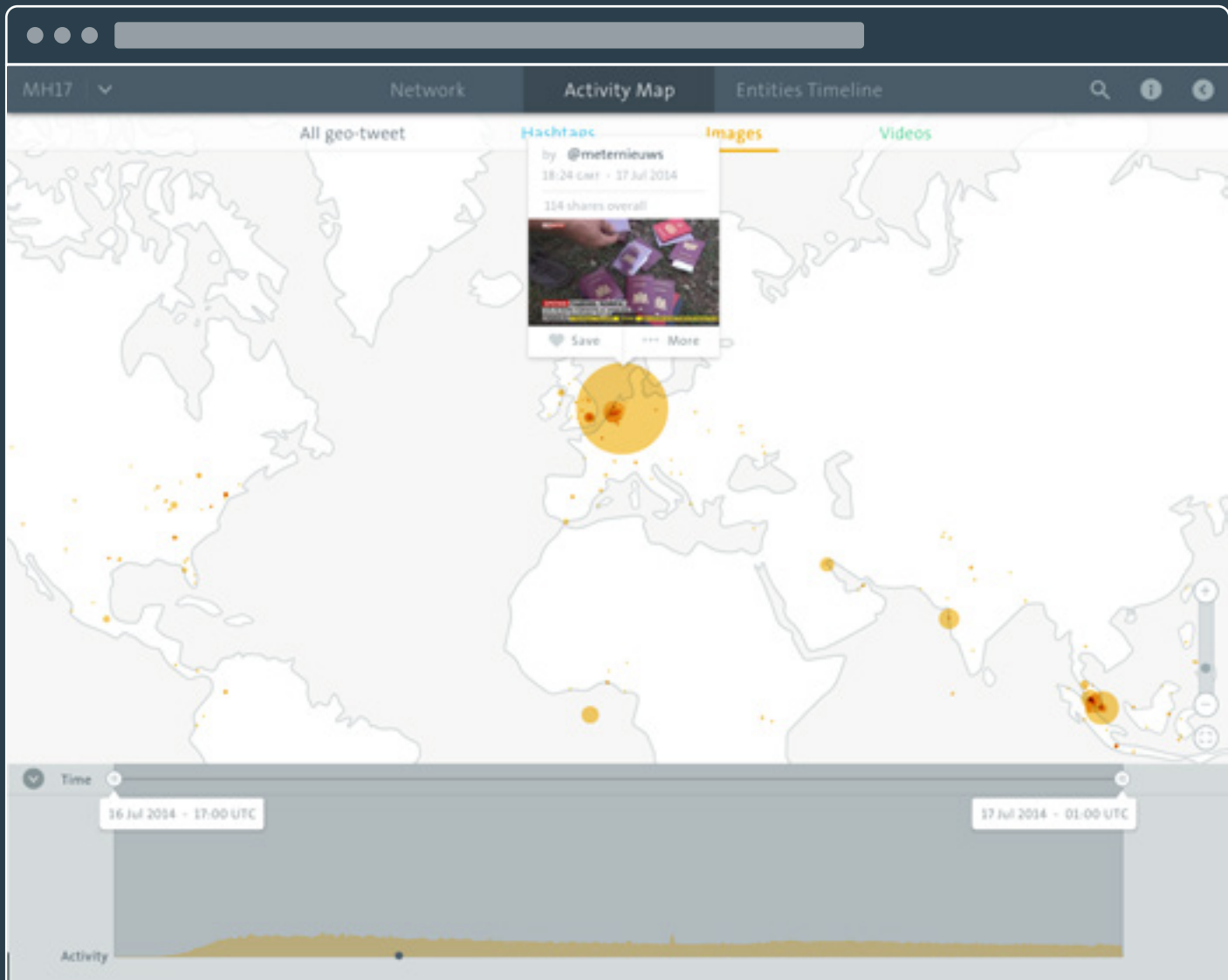
All geo-tweet Hashtags Images Videos

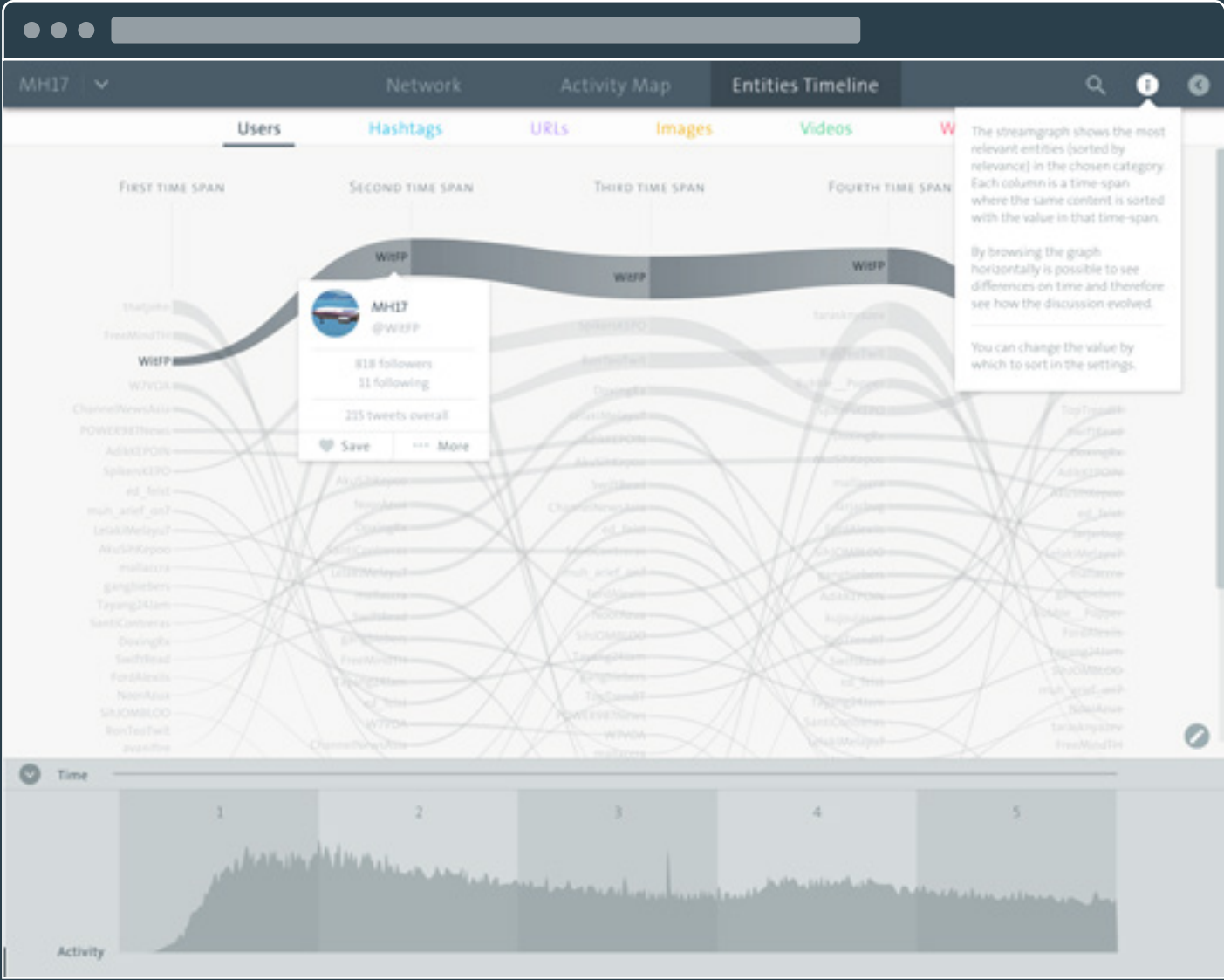
The map shows geo-localized posts.
You can look for specific locations to add a pin on the map.
Size is according to the number of retweets.

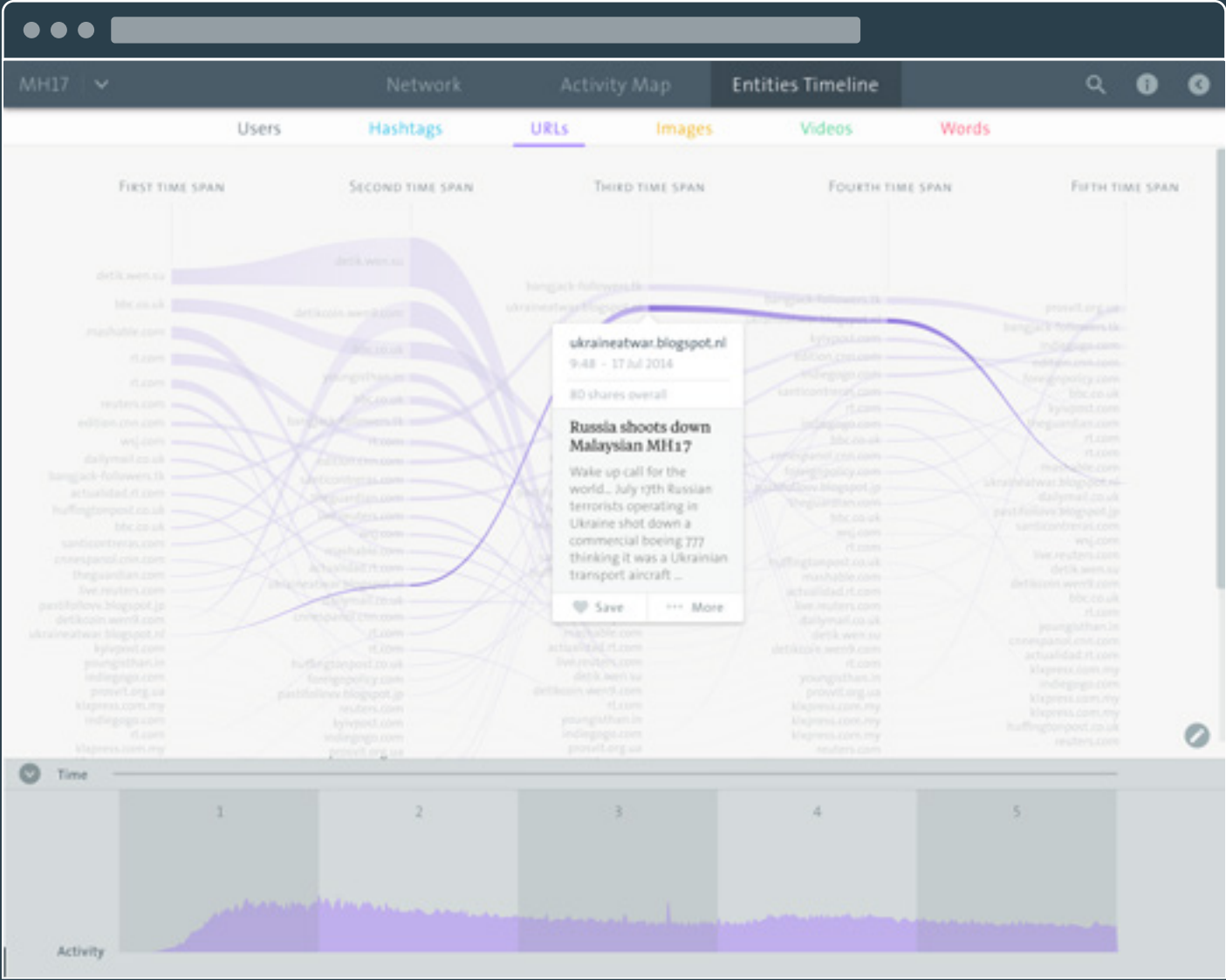
Aw AirwaysNews @AirwaysNews
16:51 UTC · 17 Jul 2014
We are following unconfirmed reports that a Malaysia Airlines Boeing 777-200, operating flight #MH370 has crashed in Eastern Europe.
132 retweets · 8 favorites

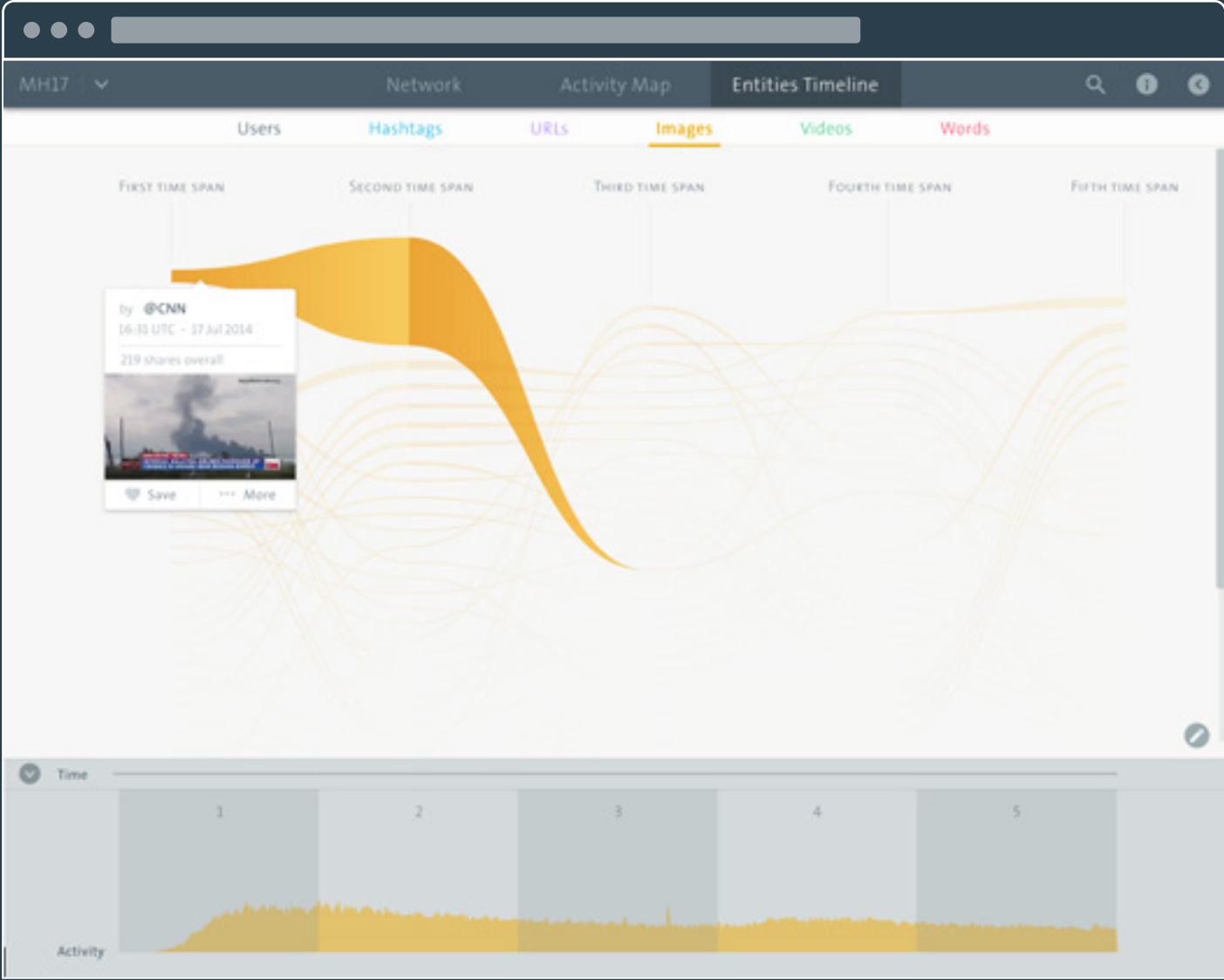
Time
17 Jul 2014 - 16:00 GMT 17 Jul 2014 - 18:00 GMT

Activity









Settings



Network

Activity Map

Entities Timeline

Project

Profile

Entities in view



Hashtags

Keywords associated with your research topic. Note that for best results the hashtag that identifies your topic research has been ignored. [MORE INFO](#) →



URLs

URLs linked together when talking about your topic.



Images

Images posted and URLs linking to Instagram and Flickr or to any image format.



Videos

URLs linking to YouTube or Vimeo or videos posted directly on the platform (e.g. Facebook or Twitter).



Users

Profiles talking about your topic within the time span selected.

Quantities

Show nodes with a value greater than:

Explanation of how the value is calculated. For example etc...
Not that a lower value means more nodes in the visualization, making it more difficult to browse and cluttered.

Appearance

Size of the nodes depends on:

Conclusion

VALIDATION

To validate the project I turned to users directly. In particular the validation has taken into account mainly two aspects of the tool: usefulness of the visual layer as a methodology to explore UGC and understandability of the visual structures adopted.

The method used for the validation consisted of an interview with Mark Vos, journalist and product manager for social publishing at Nu.nl, one of the main news portals in the Netherlands. Choosing Mark has been due to his background as journalist and for his continuous collaboration with professionals in an active newsroom (the core of Nu.nl is breaking news coverage). In addition his work is strongly connected to UGC, which he uses in many ways to improve news coverage. Hence he is a professional already into the use of social media as a news source, which helped me to focus the validation on the visualization methodology used rather than on the relevance of UGC in the workflow.

The interview was conducted by showing a rough prototype of the tool to the user and by collecting his thoughts on each designed screen. Even though it did not follow the structure of a question-answer approach, the interview has been based upon these key points:

What do you think this visualization tells?

Is everything clear?

Do you see any interaction possibility in this screen?

If so, how would you use them?

Knowing the interaction possibilities, how would those help your research?

Does this visualization provide you any insight?

Would those insights be relevant for your workflow?

In your opinion, would it be valuable for you and for others?

Can you tell me which opportunity or limitation you see in this tool?

The Network visualization appears to be confusing at first: understanding how it works and what it visualizes requires time. The user pointed out how difficult it is for strangers to data visualization to grab insights if no explanation is provided first, in particular he found unclear the reason behind nodes' position. Even if a legend is always accessible to the user, this highlights the importance to feed-forward those explanations. However, the tool is also supposed to be used with a certain frequency, which would develop in users an experience. It has to be taken in consideration that the network is a filtered and summarizing visualization, but it still represents the UGC together with all its complexity: time is needed in order to understand the key of interpretation. The good news is that time is needed only once since the visualization of a different topic or event will change appearance but not the way it is created and interpreted.

Once understood the way it works the user confirmed the expected exploratory behavior by making use of the zooming controls and by starting to look at relations between entities.

Another useful visualization is the activity of posts over time in the timeline. Of particular interest for a journalist is the presence of spikes in the chart, on which he would focus his attention to discover the reasons behind such an increment of use. Having the possibility to filter the visualization over time is confirmed to be a valuable feature in the exploration of data, especially when it comes to understanding the reasons of a change in the activity or when the interest is in the development of the conversation over time.

The activity map visualization, as known, did not represent something innovative for the user. Other tools already exist that work in the same way. The results shown on the map did not surprise the user (it is somehow obvious that the more populated areas were Netherlands and Malaysia for the MH17 case); nonetheless seeing that UK had a strong interest in the topic compare to all others regions got the user curious about it and started a fast investigation to check the reason be-

hind this exception. This proved in a simple way the effectiveness of visualization as a methodology and its usefulness in finding new perspectives on an event.

The analysis done on images in order to filter out duplicates, was confirmed to be an interesting approach to focus on original content only. This methodology makes it also more secure to rely on the content shown on the map. As thought during the design process, an additional value of this visualization is that it allows to conduct the investigation without having to use several tools. The user recognized also the usefulness of the filter options given within the visualization, such as the filter by hashtag or by kind of content. Again, the timeline component helps to understand how things developed over time, for example who were the first users in the conversation and their position when they shared their posts.

An improvement to the Activity Map would be to have a preview of the logo of the platform on which a video or an image was published from. This could help the exploration and those could become new filters for the content. The search for locations and the possibility to pin them on the map was recognized as important and expected.

The Content Timeline took the full attention of the user, curious to understand what it was about. After a brief explanation of how it works and what it shows, the user was able to understand the dynamics behind it and started to see some patterns. He agreed that such representation would indeed become more interesting when it spans a longer period of time, however the Users section already showed some directions for a further investigation.

Overall the user confirmed the validity of the tool in supporting the exploration and analysis of UGC, providing already some directions for a first check in the investigation process. Nonetheless the best use of such tool, would be of more help in the reconstruction of events and how news developed over time. Hence for breaking news reporters who deal with small real-time coverage it would not prob-

“It has potential to be a great tool for analyzing the development of news. [...] It would help to find out new angles and new ways of looking at the story.”

– Mark Vos

ably being able to provide insights with the speed required. As suggested by Mark, data journalists could be an additional target group who would benefit of this tool, especially considering their deep understanding not only of visual representation, but also on how data are gathered and filtered.

Mark pointed out a potential flaw of using one tool for the exploration: if there is a mistake in the data it will present itself throughout all the three visualizations, making the risk to not see it higher.

LOOKING BACK TO LOOK FORWARD

Working on this project has been first of all a rich experience, full of challenges and new opportunities.

The goal of the project, support the exploration of UGC by the use of a visual layer, has been accomplished with positive results, confirmed by a validation with hypothetical final users. Furthermore, during the development of the project I encountered positive and interested feedback from many professionals in the journalism field, meaning that social media content and its analysis represent a challenge of today.

Even though the gathered Twitter data was collected from a relatively small amount of time, the three visualizations in all their variables were able to give insights about how information has spread on the platform. In particular one main finding is that a previous knowledge of the topic is not required in order to understand the content of the visualization: it is instead through the exploration of the visualization itself that new knowledge can be acquired.

As confirmed during the validation, the project demonstrates to be a valuable tool for journalists and academics in news reconstruction and social media analysis, especially for its capacity to give both a macroscopic and microscopic view enhanced by a time filter.

The visual layer showed to have the potential to provide new angles and perspectives about a story, always an important element in journalism.

Further developments would touch five points:

- Expand the data gathering to other platforms, in particular Facebook and Instagram, which along with Twitter contain the most important UGC. In such case there would be need to explore new ways to combine data coming from the different platforms without the need to analyze them as separate datasets. Therefore it would be answered an important need that any kind of social media analysis has to face, observing content and relations regardless the platform on which they were published.

- In order to automatize the whole *data – visual structure* process it will be required to test the best way to filter out irrelevant entities to output the visualization as clean as possible. This has to consider the removal of spam users, content and websites based on a list of known keywords. More sophisticated filters will need advanced algorithms to determine the relevance of an entity within the context;
- The tool and the methodologies adopted are open to adjust to new kind of representations, especially in the direction of real-time visualizations in order to better serve breaking news reporters;
- The verification of UGC has been only touched by this project, since it can rely only upon complex algorithms. However a visual feedback for trustiness could easily become an implemented feature. Other platforms dedicated to qualitative debunking of rumors already exist, such as *Emergent* by Craig Silverman. After describing my project to Craig, he said their API would indeed be a valid use in the tool.
- A last thought goes to the opportunities that the new media landscape provides with the technology and variety of devices available. Even though such kind of complexity representations would simply not work on a smartphone screen, the option to create an app where the saved items will be collected can still be considered, as an extended version of the Dossier sidebar. This way the insights the user retrieved from the use of the tool will stay with him even on-the-go. Going in the opposite direction, the usage of big size screens, such as touch tables, would help in the exploration of the canvas (especially for the Network visualization).

All the mentioned opportunities are indicators of how much the tool is just in its embryonic form while it does provide a solid starting point towards an innovative exploration of user generated content.

Bibliography

JOURNALISM AND SOCIAL MEDIA

S. D. Reese; L. Rutigliano; K. Hyun; J. Jeong. 2007. *Mapping the blogosphere: Professional and citizen-based media in the global news arena*. Journalism 8:3, 235-261. SAGE Publications.

J. Li; J. Li; J. Tang. 2007. *A flexible topic-driven framework for news exploration*. KKD '09, Proceedings.

P. Miel; R. Faris. 2008. *News and information as digital media come of age*. Berkman Center for Internet & Society at Harvard University. Cambridge, USA.

R. Weaver Lariscy; E. Johnson Avery; K. D. Sweetser; P. Howes. 2009. *An examination of the role of online social media in journalists' source mix*. Public Relations Review 35:3, 314-316. Elsevier BV.

J. Leskovec; L. Backstrom; J. Kleinberg. 2009. *Meme-tracking and the dynamics of the news cycle*. KDD 09 Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM. New York, USA.

M. Ludke (Edited by). 2009. *Let's Talk: journalism and social media*. Nieman Reports 3:4.

O. Alonso; K. Berberich; S. Bedathur; G. Weikum. 2010. *Time-based exploration of news archives*. HCIR 2010, Proceedings of the Fourth Workshop on Human-Computer Interaction and Information Retrieval.

A. Hermida. 2010. *Twittering the News*. Journalism Practice 4:3, 297-308. Routledge. London, UK.

M. Karlsson; J. Strömbäck. 2010. *Freezing the Flow of Online News*. Journalism Studies 11:1, 2-19. Routledge. London, UK.

H. Kwak; C. Lee; H. Park; S. Moon. 2010. *What is Twitter, a social network or a news media?* Proceedings of the 19th international conference on World Wide Web. ACM. New York, USA.

W. Stassen. 2010. *Your news in 140 characters: exploring the role of social media in journalism*. Global Media Journal 4:1. Stellenbosch University.

S. Waldman. 2011. *Information Needs of Communities: The Changing Media Landscape in a Broadband Age*. DIANE Publishing.

A. Bruns; J. Burgess. 2012. *Researching News Discussion on Twitter*. Journalism Studies 13:5-6, 801-814. Routledge. London, UK.

B. Franklin. 2012. *The Future of Journalism*. Journalism Studies 13:5-6, 663-681. Routledge. London, UK.

A. Hermida. 2012. *Tweets and truth: Journalism as a discipline of collaborative verification*. Journalism Practice 6:5-6, 659-668. Taylor & Francis.

E. C. Jr. Tandoc; M. Jenner. 2012. *Analyzing Web Analytics, How Newsrooms Use Web Metrics in News Construction and Why*. University of Missouri, USA.

I. Flaounas; O. Ali; T. Lansdall-Welfare; T. De Bie; N. Mosdell; J. Lewis; N. Cristianini. 2013. *Research Methods in the Age of Digital Journalism. Massive-scale automated analysis of news-content- topics, style and gender*. Digital Journalism 1:1, 102-116. Routledge. London, UK.

U. Hedman; M. Djerf-Pierre. 2013. *The Social Journalist*. Digital Journalism 1:3, 368-385. Routledge. London, UK.

A. Hermida. 2013. *#Journalism*. Digital Journalism 1:3, 295-313. Routledge. London, UK.

E. Newton. 2013. *Searchlights and Sunglasses: Field notes from the digital age of journalism*. Knight Foundation. <http://goo.gl/ShGJQq>

F. Vis. 2013. *Twitter as a Reporting Tool for Breaking News*. Digital Journalism 1:1, 27-47. Routledge. London, UK.

P. Grabowicz. 2014. *The Transition To Digital Journalism*. The Knight Digital Media Center, University of California.

U. Hedman. 2014. *J-Tweeters*. Digital Journalism 1-19. Routledge. London, UK.

R. K. Nielsen; K. C. Schröder. 2014. *The Relative Importance of Social Media for Accessing, Finding, and Engaging with News*. Digital Journalism 2:4, 472-489. Routledge. London, UK.

S. Schifferes; N. Newman; N. Thurman; D. Corney; A. Göker; C. Martin. 2014. *Identifying and Verifying News through Social Media*. Digital Journalism 2:3, 406-418. Routledge. London, UK.

C. Silverman (Edited by). 2014. *Verification Handbook, A definitive guide to verifying digital content for emergency coverage*. European Journalism Centre. Maastricht, The Netherlands.

DESIGN AND VISUALIZATION

E. R. Tufte. 1990. *Envisioning Information*. Graphics Press. Cheshire, US.

G. Anceschi. 1993. *Il Progetto delle interfacce: oggetti colloquiali e protesi virtuali*. Domus Academy.

S. K. Card; J. D. Mackinlay; B. Shneiderman. 1999. *Readings in information visualization: using vision to think*. Morgan Kaufmann Publishers. San Francisco, US.

J. E. Swan II; T. Rhyne; D. H. Laidlaw; T. Munzner; V. Interrante. 1999. *Visualization needs more visual design!* IEEE Visualization, 485-490.

G. Bonsiepe. 2000. *Design as tool for cognitive metabolism: From knowledge production to knowledge presentation*. International Symposium "Ricerca+ Design" at Politecnico di Milano, Italy.

E. R. Tufte. 2001. *The Visual Display of Quantitative Information* (2nd Edition); Graphics Press. Cheshire, US.

C. M. Dal Sasso Freitas; P. R. G. Luzzardi; R. A. Cava; M. A. A. Winckler; M. S. Pimenta; L. P. Nedel. 2002. *Evaluating Usability of Information Visualization Techniques*. Proceedings of IHC Fifth Workshop on Human Factors in Computer Systems, 40-51.

G. Bellinger; D. Castro; A. Mills. 2004. *Data, information, knowledge, and wisdom*.

<http://www.systems-thinking.org/dikw/dikw.htm>

C. Ware. 2004. *Information Visualization. Perception for design*. Morgan Kaufmann Publishers. San Francisco, US.

B. Craft; P. Cairns. 2005. *Beyond guidelines: what can we learn from the visual information seeking mantra?* Proceedings of the Ninth International Conference of Information Visualisation, 110-118. IEEE.

J. Abrams; P. Hall. 2006. *Else/where: Mapping New Cartographies of Networks and Territories*. University of Minnesota Design Institute.

A. Lau; A. Vande Moere. 2007. *Towards a model of information aesthetics in information visualization*. 11th International Conference on Information Visualization, 87-92.

G. Scagnetti; D. Ricci; G. Baule; P. Ciuccarelli. 2007. *Reshaping communication design tools*. IASDR07 Emerging Trends in Design Research.

J. Yi; Y. Kang; J. Stasko; J. Jacko. 2007. *Toward a deeper understanding of the role of interaction in Information Visualization*. IEEE Transactions on Visualization and Computer Graphics. 13:6, 1224-1231.

L. Manovich. 2008. *Introduction to Info-Aesthetics*. <http://goo.gl/POAqB9>

A. Burdick. 2009. *Design without Designers*. Keynote for a conference on the future of art and design education in the 21st century. University of Brighton, England.

L. Masud; F. Valsecchi; P. Ciuccarelli; D. Ricci; G. Caviglia. 2010. *From Data to Knowledge. Visualizations as Transformation Processes within the Data-Information-Knowledge Continuum*. Information Visualisation IV. 14th International Conference 2010, 445-449.

P. Ciuccarelli. 2012. *Visual Explorations. On-line Investigations for Understanding Society*. In J. Errea; A. Gil. Molofiej 19th International Infographics Awards. 6- 23.

B. Latour; P. Jensen; T. Venturini; S. Grauwin; D. Boullier. 2012. *'The whole is always smaller than its parts' – a digital test of Gabriel Tarde's monads*. *The British Journal of Sociology* 63:4. London School of Economics and Political Science.

R. Rogers. 2013. *Digital Methods*. MIT Press. Cambridge, US.

G. Ubaldi; G. Caviglia; N. Coleman; S. Heymann; G. Mantegari; P. Ciuccarelli. 2013. *Knot: an interface for the study of social networks in the humanities*. Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI.