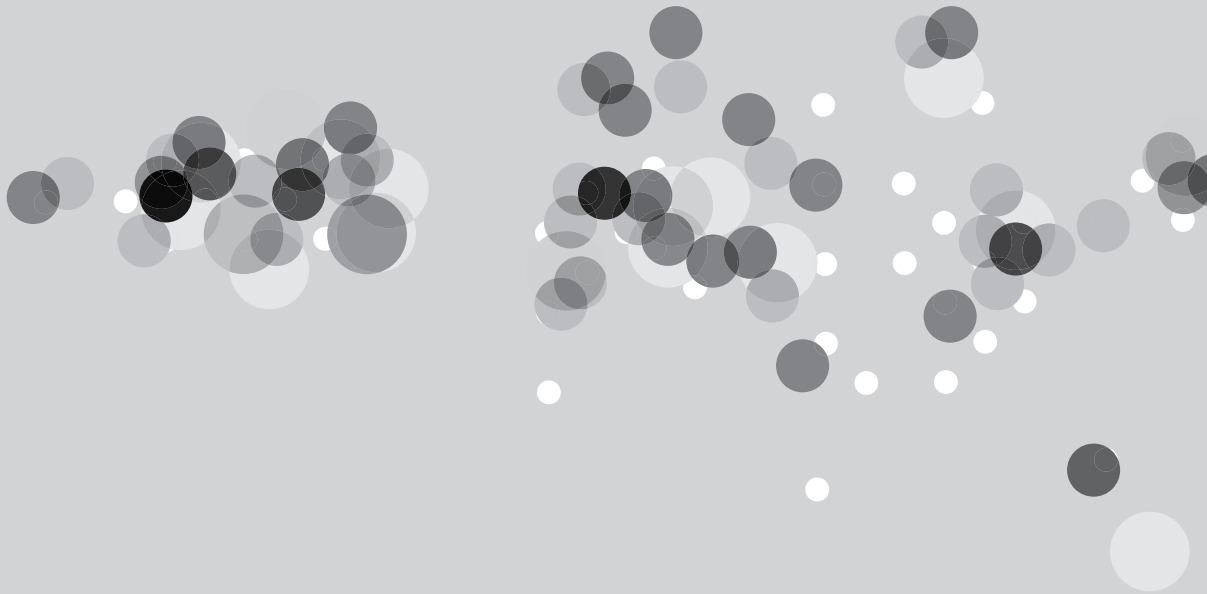
01. Full papers

2 : Interactive Data Visualization



Form follows what? Some Thoughts about Behaviours and Data Architecture in Dynamic Data Visualization

Alfredo Calosci

negot.net

alfredoc@negot.net

Abstract. The aims of this paper is to underline the key role played by the way we "shape" data architecture especially when we try to visualize and communicate a complex process.

Keywords: Information Architecture / Ontologies / Narrative)

1 Introduction

What makes the difference between Interactive Data Visualization and the others two main topics of this "2CO Communicating Complexity Conference" - Informative Animation and InfoGraphics - it's real time and the ability of dealing with a huge amount of data.

Most of the times, Interactive Data Visualizations allows real time interaction for data manipulation (sorting, filtering, browsing ...), sometime they are dynamically built on real time data, in some occasion they show both aspects.

In most of these applications, the visual aspect of the "signifiers" is defined by a set of rules / behaviors that "translate" some aspects of the data architecture into graphics attributes (shape, position, color and so on).

This is actually the way all "data visualizations" have always been done; the only new thing is that, in a digital interactive application, we are now forced to describe those rules into an algorithm.

What users will see in our applications is the result of the combination between our visual algorithm and the data structure used to describe the process.

2 Why we visualize

We visualize because thinks are sometimes easier to understand (or to remember) in the visual - illustrated - format rather than as a text or a table of figures.

Cognitive sciences are there to tell us why but let's start considering the following examples:

Case 0: reverse visualization

We probably have all learned in our algebra classes that:
$(a+b)^2 = a^2 + b^2 + 2ab$

The formula above is not a very complex one but it's much easier to remember if we know what it represent.

But we have to admit that this example is tricky, the formula is a "compact" way of representing the same geometry showed in the illustration.

The abstract algebraic format may be harder to understand but it's easier to manipulate in others context.

At a certain point in human history it was an achievement to be able to represent geometric properties in a symbolic way rather than by drawing.

This case could be considered as an example of "reverse" visualization".

As many authors pointed out, we should never forget that texts, notations and images are all part of the same communication system [1]; Shapes and colors and lines, without a shared semantic agreement are not able to communicate by themselves. [2]

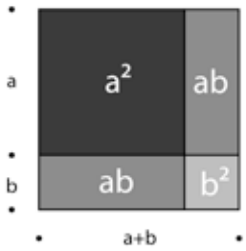


Fig. 1. (a+b)2 = a2 + b2 + 2ab

Case 1: dissemination: to translate into visual language

Not all the semantic agreements are equally easy to learn neither are they universally understood.

We still consider the language of images a more direct and universal one. Icons and visual metaphors have proved over the history their unique capacity to bridge linguistic and cultural boundaries.

So what makes experiences like ISOTYPE so different from, let's say, a romanic cycles of painting ? It is not, of course, the attempt to create a visual narrative for educational porpoises; it's rather the challenge to conceive a whole new communication system with its own grammar and syntax elements.

If we take a look at the classic ISOTYPE workflow, we will find on one side the editor (the curator), selecting the topic to be illustrated and the data to be showed.

On the other side we have the artist, who create the elements of the well none flat iconic alphabet.

And finally, in the middle, we have the transformer, the one who apply the ISOTYPE rules and give shape to the final illustration, searching for the best combination of elements to express the communication goal.

Now, this is exactly the role of nowadays visual algorithms, the ones we need to introduce real time data and/or interactive manipulation as part of this visual narrative.

The ISOTYPE method produced many beautiful illustrations, translating facts and figures into compelling visual metaphors, but very rarely suggest a "cause > effect" relation between the variables on display.

Correlation and causation, even if the first does not imply the other, are what statistic and scientific visualizations are all about.

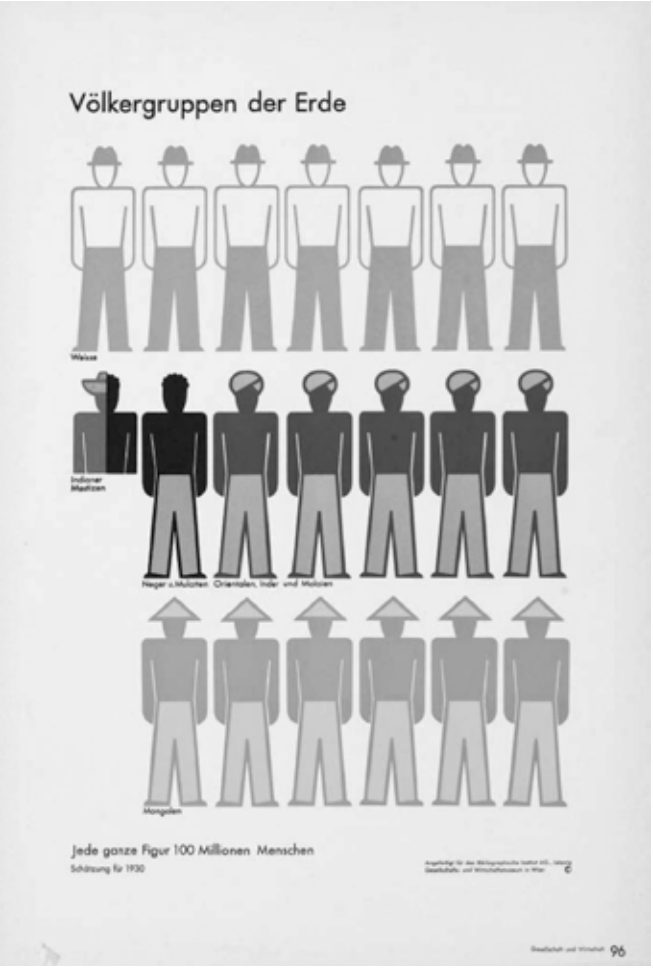


Fig. 2. Völkergruppen der Erde. - Gesellschaft und Wirtschaft

2.3 Case 2: analyze and discover, the statistic/scientific approach

Plot your data in a meaningful way and you may highlight the correlation between your variables.

It does not matter if you are testing your hypothesis, communicating your vision of a process, monitoring a system, or trying to discover something new.

John Snow's Cholera Map of London, in 1854, is the paradigm of many, sometimes much more sophisticated, diagrams used in statistic and, in broad terms, in science.

What all those diagrams have in common is not the lack of any kind of decoration, as Edward Tufte would suggest, but it's rather the attempt to describe a process showing the relations between its variables.

Scientific data visualization can afford some amount of illustration without any decrease in performance, it's just a matter of cultural / esthetic options; none other than ISOTYPE taste for flat icons and bi-dimensional diagrams.

Both approaches can lead to "dumb" visualizations when they just display a bunch of unrelated figures dressed with in an info-graphic esthetic.

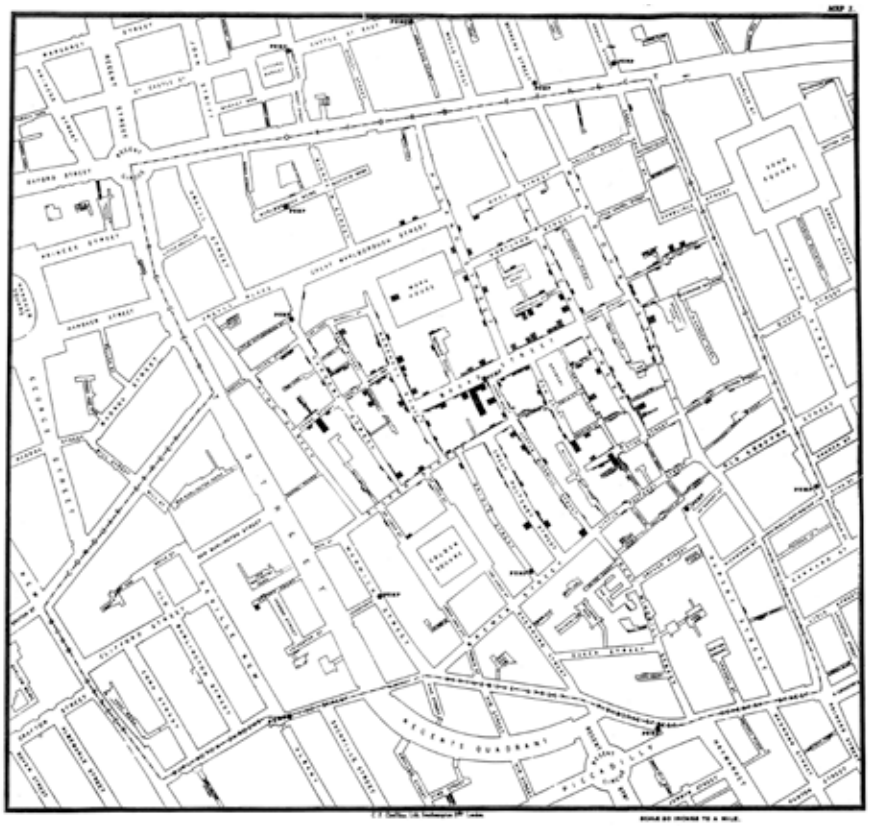


Fig. 3. John Snow – London 1854

3 Data wildlife - real time data in their natural environment

Real time data are often hard to hunt, even if thinks are now getting easier. If we look at the data visualization workflow proposed by Ben Fry [3]:
"Acquire > Parse > Filter > Mine > Represent > Refine > Interact"
more than half of the process has to do with how to get the data and how to make them work properly.

3.1 the post-desktop scenario

One of the most friendly environment to go for data is the "post-desktop" scenario created by cloud computing + web application and their respective APIs.

In this case data come in a very structured way and it's quiet easy to filter the kind of information we are looking for.

Flickr, Twitter ... but also other local "smart services" like bicycle renting network and so on, allows others to build their own "side" applications providing access to the same core data.

We all know so many examples fueled by data coming from those platforms but, among them, we would like to mention "Emoto London 2012", a project created by Moritz Stefaner, Drew Hemment, and Studio NAND with the aims of tracking the emotional response to London Olympics while they where still in progress. [4]

According to the authors: "... We track twitter for themes related to the Games, analyze the messages for content and emotional expressions, and visualize topics and tone of the conversation".

The result was a kind of a thermometer, shaped as a sophisticated origami-like diagram, ranging from "extremely negative" to "enthusiastic".

This was made possible by the "extra" algorithm that assign an emotional value to each twit; it could be a way to explore: 99% of API's structured data plus a few line of extra smart code.

3.2 semantic web and open data initiatives

The WWW have been semantic since the very beginning but most of the content is still "unlabeled".

Find the right schema for our contents, and being consistent in applying those rules every time we publish something new, is not a task for lazy people but it should be a must for every OpenData initiative.

Public data without a structured framework are better than nothing but they may result useless.

As an anecdote, the spanish house of representatives published for the first time in 2011 the "declaration of income" of each elected as a raster image of the income form, in pdf format.

A public crowdsourcing call, made by the association "Pro Bono Público", allows to transform all those almost unusable bytes into a proper database [5]. People was asked to "adopt" an elected and type the income form figures into a shared Google-Doc spreadsheet [6]. The task was completed in less than 4 days.

3.3 mining for data

Where existing structured data fails or when they are not available is the only alternative to explore.

3.4 your own data

Here is where we get to the crucial point; our opinion is that the quality of any data visualization rely mainly on the way we shape our informative architecture. By adding some smart features to data coming from an API, by crossing datasets coded in an "exchangeable" format, by mining the web or adopting the right standard for describing every variable we need to describe a certain process.

Let`s have a look at this "linear" example. Last year I was asked to create a 4 meter by 4 relations diagram for a public exhibit. [7]

The initiative, called "Territorio Archivo" and promoted by "Fundación Cerezales Antonino y Cinia",was about "collecting memories" among neighbors from 6 small villages close to Leon, in the north of Spain.

The structure of the graph is quite simple: a little bit more than 800 pictures connected to their owners (around 50 people) connected to their home village.

The goal was to give some kind of "synoptic" view of the photographic archive collected during the initiative.

A simple "force directed" algorithm was enough to produce the final result. Whats happened is that the "natural" configuration of the forces tended to put the biggest village in the center and this was considered *politically incorrect* by the curator who preferred a more *balanced* representation.

At the end the illustration was "manipulated" according to the curator wish for a non centric diagram.

But the key point here is that it doesn't matter what kind of algorithm you run; the more connected to the center, the heavier to the bottom, the biggest on the top … the central role of Cerezales del Condado, the village where most of the pictures where collected, would emerge anyway because it's part the dataset and relations are shaped into the data architecture NOT in the visual algorithm.

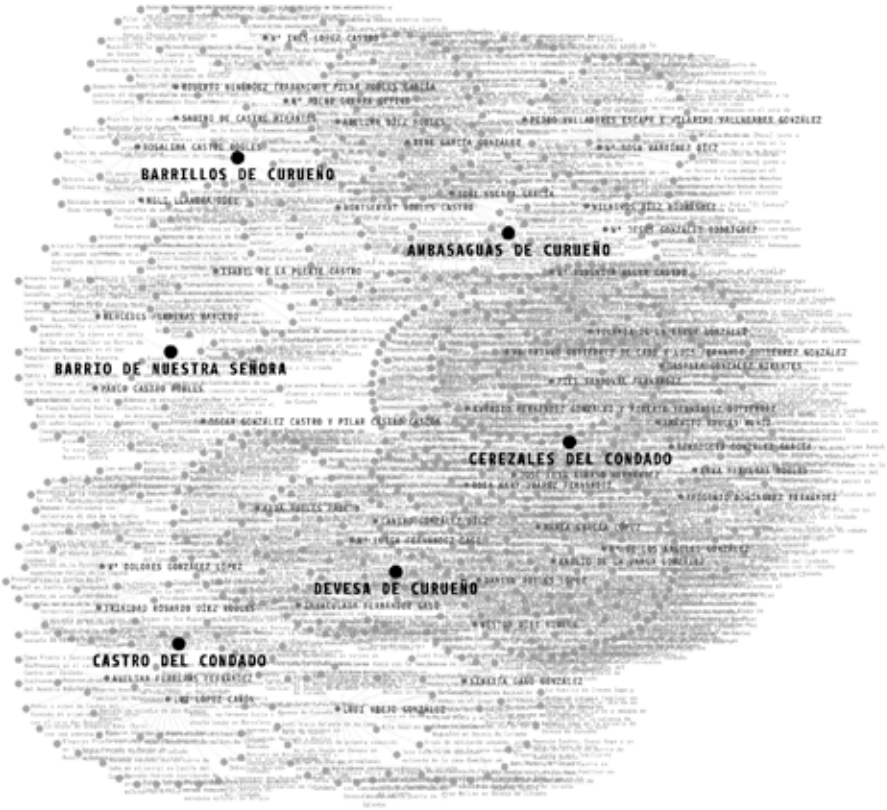


Fig. 4. Territorio Archivo – a force directed diagram.

4 Facing complexity

This is especially true when we have to deal with complex process.

Complexity is not difficult to manage, it's unpredictable. Most of Generative Art algorithms are able to create a complex / chaotic behaviour with a few line of code and starting from very simple rules. From an artist point of view, coding is a way to "loose control" over the final result, it's about creating a process and let him work.

In design, most of the times, we have a communication goal to achieve (even if loosing control over the result may still be a way to accomplish it). Complex processes are often difficult to explain because the cause > effect relation is no longer visible and this is because of the feedback.

Feedback is the "trademark" of complexity; the properties of a complex system "emerges" from the relations between its parts and they are no longer there when you split the system into its components.

In order to face complexity we have to put the focus of our attention to relations, rather than objects, to qualities instead of quantities and to patterns more than on elements.

For this reason, a semantic "labeling" of our digital content may not be enough if our goal is to describe the quality of those relations and the system as a whole (who is greater than the sum of its parts) as the the main future of our information architecture.

Conceptual Reference Models (CRM) and others kinds of Data Ontology may represent a more promising approach. Let's consider the CIDOC-CRM (ISO 21127:2006), developed by the International Council of Museums (ICOM) with the aim of: " … promote a shared understanding of cultural heritage information by providing a common and extensible semantic framework that any cultural heritage information can be mapped to". [8]

CIDOC-CRM can be viewed as a predefined catalog of Entities and Properties.

Almost half of the Entities are "Temporal entities" like Periods or Events, the other half describes "Persistent items" like Things or Actors, and the whole system is completed with other less articulated entities like Places or Dimensions.

Properties, that may be assigned to Entities, are in fact, relations like "took place at", "is composed of", "is carried by" or "was produced by".

This paper, for example, could be described as a: CRM Entity (E1) a Persistent Item (E77) a Conceptual Object (E28) carried by (P128) a Physical Man Made Object (E24) … and so on … filling up entities and properties like the author's name or the dimensions of the proceeding book.

Without getting into further details, there are two aspects we would like to underline, first: CIDOC-CRM is time / event based, it allows to describe "objects" all along their life, from the conception, to the production, to the conservation or others aspects of cultural heritage management. Second: it's not a dataset itself, it's rather and attempt to allows interoperability among existing data.

To consolidate datasets belonging to different cultural entities is not an easy task neither from the technical or the conceptual point of view. This kinds of ontologies may represent a solution for the conceptual part.

A good map of entities and relations is probably the best starting point for any attempt to explain / visualize complex processes.

You may not need it if your communication strategy is based on linear media, as animation, and your work is to translate the dynamic of a complex system into a story.

In other interactive visualization strategies, like "simulators" or "complexity browsers", that allows users to discover by themselves complex behaviors and the appearence of emergencies, all the performances have to be well shaped into the data architecture.

5 Conclusions

Narrative, in the broad sense, is as a sequence of cause > reactions facts (acts) that take place in time and space.

On the other hand, the mind and our cognitive system seams to be driven by "action", we explore our environment looking for affordances for our planned or potential activity. Doing so we tend to create interpretations of what happens around us as a chain of cause > reactions sequences of facts.

Even our memory work quiet well remembering stories. Many memory techniques, like the "journey method", uses sequences of places to help us storing information.In this mnemonic system … "The items to be remembered are mentally associated with specific physical locations. The system relies on memorized spatial relationships to establish, order and recollect memorial content." We still say "In first place …" when we have to point out our opinion on a certain subject.

We may say that Knowledge itself may be, at the and, based on stories.

Now, if we look at some of the most promising data ontologies, like CIDOC-CRM we realize that they are events / time based.

They don't just describe how things are but rather whats happened to them, they focus on the process that bring then to life and transform them over the time.

Is the connection between storing data and telling stories much stronger than we supposed ?

Notes

[1] ELKINS, James; "The Domain of Images"
[2] FALCINELLI, Riccardo; "Guardare, Pensare, Progettare, Neuroscienze per il Design"
[3] FRY, Ben; "Visualizing Data: Exploring and Explaining Data with the Processing Environment
[4] Emoto: Visualising the Emotional Response to LONDON 2012
http://www.emoto2012.org/
[5] Pro Bono Público: adopta un senador
http://blog.probp.org/post/17031929224/transparencia-con-una-hoja-de-calculo-adoptaunsenador
[6] adopta un senador.
https://docs.google.com/spreadsheet/ccc?key=0AowzHU9kHzeudG9aSjVVOEQxVHpzR2E4ZDdhVXJLQlE&hl=en_US
[7] Territorio Archivo:
http://www.fundacioncerezalesantoninoycinia.org/territorioarchivo/
[8] CIDOC-CRM http://www.cidoc-crm.org

References

CAPRA, Fritjof "The Web of Life, A New Scientific Understanding of Living Systems" - New York: Anchor Books, 1996

ELKINS, James; "The Domain of Images" - Cornell University Press - 2001

FALCINELLI, Riccardo; "Guardare, Pensare, Progettare, Neuroscienze per il Design" - Stamapa Alternativa & Graffiti / Nuovi equilibri - Viterbo 2011

FRY, Ben; "Visualizing Data: Exploring and Explaining Data with the Processing Environment", O'Reilly Media - Cambridge, Massachusetts USA - 2007

NEURATH, Marie, KINROSS Robin, "The Transformer: Principles of Making Isotype Charts", Hyphen Press - London 2009

Interaction Design and Data Visualization for Future Digital Cars

Andrea Di Salvo, Paolo Marco Tamborrini, Fabrizio Valpreda, Andrea Gaiardo

Polytechnic of Turin, DAD, Viale Mattioli 39, Turin, Italy

andrea.disalvo@polito.it, paolo.tamborrini@polito.it,
fabrizio.valpreda@polito.it, andrea.gaiardo@polito.it

Abstract. Driving a car is becoming more and more complex. This does not happen because the functions that are necessary to drive are more difficult to perform: on the contrary, for example, electric and hybrid vehicles do not have any manual transmission gear. However, complexity lies in the fact that cars shall provide information on their current status, projections on their future state, optional information related to infotainment, to points of interest for the driver or the passengers, to the mobility system, to the connection with other subjects and objects outside the car etc.
In such a variegated scenario, it is necessary to design a new space and several types of visual and interactive languages able to show and guarantee both pieces of information that are necessary to drive a vehicle in a safe way, and the accessory ones. Space and information shall then be designed in order to be adaptable to the user. The paradigm of adaptability has often been confused with customization. In this case, it aims instead to provide an intelligent car-system able to understand both the attitudes and the needs of the driver and to provide real-time answers, as appropriate to the situation.
Modern cars are characterized by the constant effort to prove their potential in terms of performance and new integrated accessories, such as the protection from external agents, impacts and noise; to this end, visual languages can conform to this trend or they can express new forms and representations of data through which they can describe an increased need for communication with the external environment, a new awareness in the use of the car or a desire to focus only on driving etc.
Moreover data visualization is changing thanks to a new mode of interaction between man and commands, as signs and their directionality cannot be inconsistent with the new input mode based, for example, on the use of gestures. In addition, on one hand, the introduction of gestures and other interaction modes inside the car can be a way to find a solution to the growing presence of controls and buttons which can often distract, especially if they are operated through a touch screen; on the other hand, it could become part of the driving pleasure, something that today is more and more difficult to feel. In this way, data visualization and interaction design can cooperate to provide a new sensory experience to the user.
The research lead by the HMI team of the Polytechnic of Turin in collaboration with CRF seeks precisely to give solutions to these challenges in a Human Centered Design approach.

Keywords: Interaction design / Sustainable Mobility / Visual Languages

1 Introduction

Technology applied to the automotive sector allows to insert a huge amount of data and information inside a car. The ability to manage this flow and make it accessible to both the driver and the passengers is one of the most complex research areas of interaction design.

In particular, the fact that the interior compartment of a car is now packed with several digital screens is raising many issues in terms of usability, and specifically in terms of interaction and visualization. It is possible indeed to interact through complex tools, like multifunctional rotary knobs, or through touch screens, which however require observation. Moreover, information displayed is often redundant compared with actual needs, it is located in separate areas, distant from each other and - which is more concerning - far from the user's focus of attention: the road.

This can often cause an information overload having an impact on the perception of the situation, the mental workload and the knowledge of the car status and the surrounding environment. [1]

Despite these undeniable problems, technological skills should not be limited to the mere visualization of indicators and gauges that are strictly necessary for driving. On the contrary, they can be used to show useful information according to the users, the surrounding environment and the current situation, through the identification of the correct methods and timing. [2]

Among its multiple goals, the research project aims to design a graphical interface concept and study its interaction inside the interior compartment of an EV. The actual result of the research is presented in this paper: it consists of a hierarchically structured information map, which is adaptable according to the driver's requirements, the situation and the context. The adaptability can also be managed, automatically by the vehicle storing daily data and driver's demands, or consciously by the driver. Compliance to strict and rigid safety margins will be guaranteed during the whole work. The project, currently in the prototyping phase, needs user's tests and feedbacks which are not yet available.

2 Scenario

Despite the different type of engine, EV formally maintain the same characteristics of a vehicle with combustion engine, both externally and internally. The current structure still stems from the morphology adopted and consolidated in the 1920s. Nevertheless, if cars can, or perhaps have to, interconnect the vehicle and the user with the external world and with other users simultaneously present in the street , the information they have to show is very different, both qualitatively and typologically. In fact, while it is true that information regarding battery life can be compared to the one related to remaining fuel, implications are very different. Today the autonomy of an EV is very limited, charging stations are far fewer than normal gas stations and there is a huge difference between charging

and refuelling time: a few minutes in the latter case, up to several hours in the first one. This example shows that additional support information is required in an EV, such as the distance from the nearest station and possible alternatives, the availability of charging points, energy cost and quality, and maybe also information on the energy source which can be more or less sustainable. For these reasons, EV can represent an excellent research area, given the quantity and complexity of available information and the time needed to display them.

EV should also be considered in the framework of a complex mobility scenario in which the relation not only between different means and modes of transportation, but also with infrastructures, political support or opposition, emotional factors affecting the daily choice of a means of transportation and the geographical and cultural context should be taken into consideration.

The main actors of the system and their relationships, according to which the driver is placed at the centre of the system, are shown in Figure 1 [3].

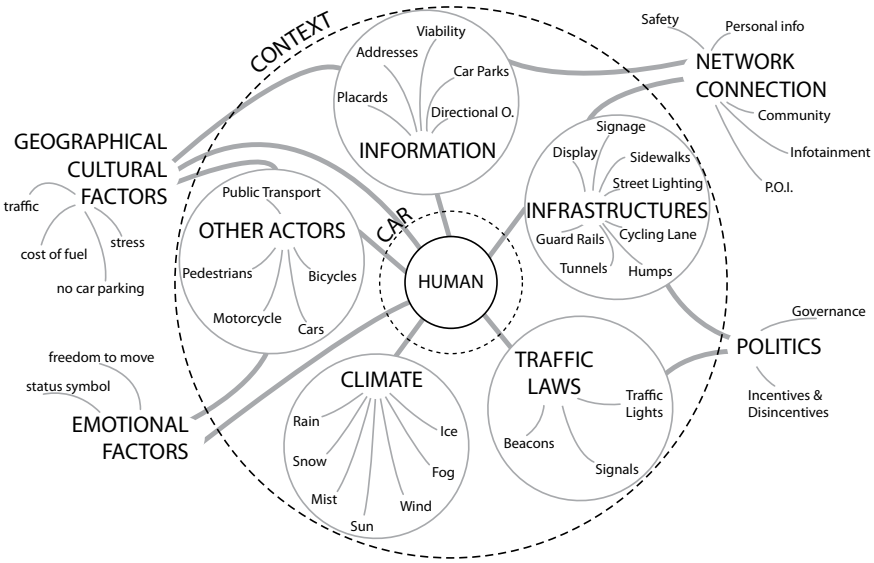


Fig. 1. The main actors of the mobility system and their relationships

Until a few years ago, cars could provide information only on the state of the vehicle, often not in an exhaustive way. Today, thanks to the use of sensors, cameras and the ability to connect to infrastructures V2I, other vehicles V2V, networks or other systems V2X, many factors can be detected and brought to the driver's attention through visual interfaces. Cars are then proposed not only as protective shells but also as cognitive and perceptive filters. The filter in some cases can act as a bottleneck, filtering and selecting only some of the incoming data; in other cases, it can increase the perception and the awareness of the situation, providing real-time data.

The massive introduction of technological devices into electric and hybrid cars has transformed these products into some kind of laboratory where to experiment new possible functions. However, there are several attempts of integration with the car: the complete implementation of car-system devices or the connection of user's nomadic devices, able to communicate with the vehicle system. Applications using the camera of the device to provide information on, for example, braking distance and lane departures, are already available on Google and Apple online stores. A solution does not exclude the other. The main differences lie in use, in computing capacity - which today is higher in smartphones, compared to on-board computers - in data storage, in portability, in integrated components and in the ability to continuously update the software.

Fig. 2. iOnRoad App Screenshot

3 Methodology

The adopted methodology combines the design culture [4], the interaction design [5] and the human centred design [6]. The first phase of the analysis deals with the critical comparison of dashboards and their interaction tools currently available on the market, widening the scope of investigation to also include, for example, concepts presented during exhibitions and workshops. First of all, a study of interfaces currently appearing on electric and hybrid vehicles has been conducted. On one hand, the decision to consider also hybrid vehicles has been taken to enjoy the opportunity to analyze a wider number of case studies and, on the other hand, because dual power supply increases the level of complexity, with many implications. The critical analysis focuses mainly on two aspects: highlighting similar communication structures and identifying the relationships between different types of information. At this stage, the investigation does not critically analyze the symbol itself but it limits to the information conveyed by

the symbol, without going into stylistic or aesthetic issues concerning, for example, the line including the dashboard.

As shown in Figure 2, the information is organized in very different ways, but it can be summarized into three categories: symmetrical models with respect to the y axis with the speedometer at the centre, models based on concentric circles and models that divide the information into two reading levels, using HUD or similar systems.

One of the highlighted problems is the lack of clear connections between the driver 's behaviour, the vehicle status and the surrounding environment. The attempts to communicate more or less virtuous driving styles are, in fact, often disconnected from the system and they use metaphors taken from the greenwashing language, for example a plant that grows or decreases, or a ball that has to be kept in balance.

Another weakness is the crowding of too much information, which generates confusion in users. It should also be emphasized that in most cases the infotainment information are positioned at the centre of the dashboard and are editable through a touchscreen. If Norman stresses the difficulty of interaction with hard buttons, also due to parallax problems, in this case we believe that without the aid of physical constraints, such difficulties are exacerbated.

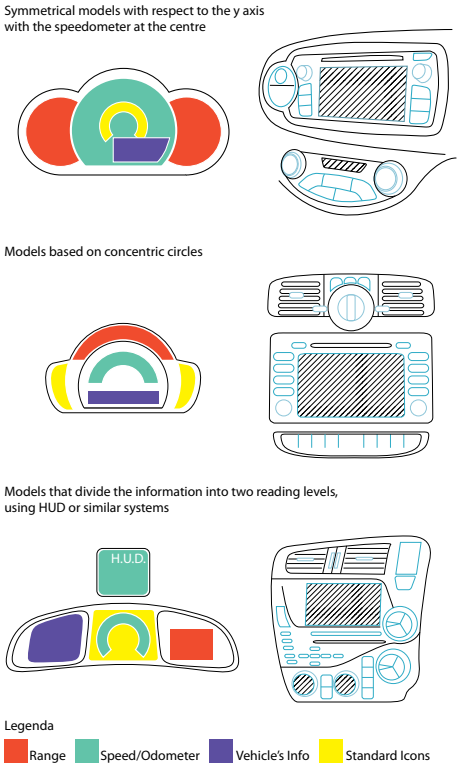


Fig. 3. Critical comparison of dashboards and their interaction tools

SmartGauge [7], designed for Ford by Ford Motors, Smart Design and IDEO, represents a case study focusing on adaptable interfaces for EV. The interface is designed on the basis of studies on actual readability and on visual and cognitive impact, around a digital speedometer from which, according to the user's request, several windows containing data and indications appear, helping the driver to focus on his or her driving style in order to reduce fuel consumption. In this case, information can contract or expand and this solves the problem of cognitive load in a simple way. However, they do not show any data different from the ones that are usually shown.

Fig. 4. The SmartGauge Interface

Usually, information regarding navigation or infotainment is placed on a central panel, also accessible to passengers. In recent years, perhaps to adapt to smartphones interaction modes, these panels have often been operated through touch screens. In the case of complex and branched menus, instead, rotary commands are preferred. They can be rotated to scroll through the different functions in the menu, and they can be pressed to select the desired function, as with the following devices: BMW iDrive[8], the Audi MMI System [9] and the Mercedes COMAND systems [10].

Several studies have suggested that these commands, despite appearing very simple in their external aspect, are very difficult to use because of the menu structure itself [11].

The research continues with an information categorization.

Unlike Tönnis et al. [12], who draw up a hierarchy of tasks, in this case the information itself is divided into primary, secondary and tertiary information. Primary information is always visible and necessary while driving, also in full compliance with current regulations. Secondary information can be displayed on demand while driving and includes, for example, navigation and parts of the infotainment system. Finally, tertiary information is the one retrieved when the vehicle is not moving or when the engine is switched off.

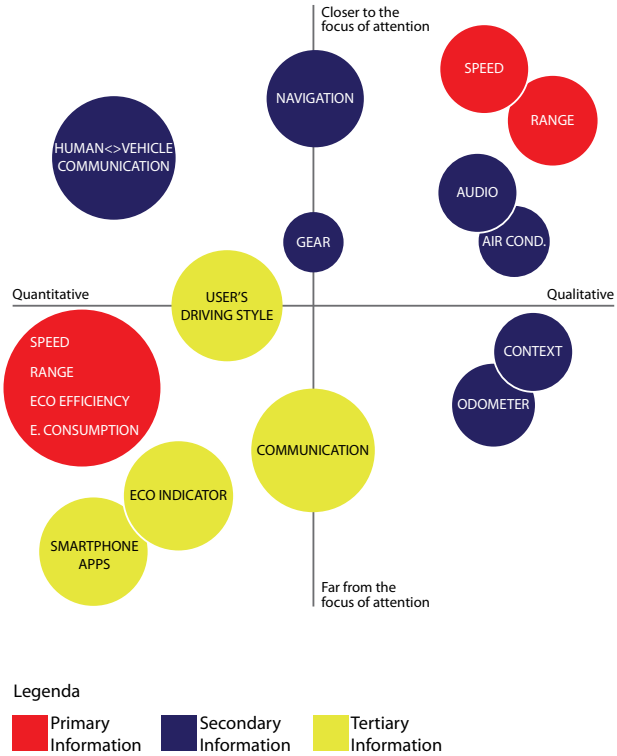


Fig. 5. Hierarchy of Information

This hierarchy has to take into account a rapidly-expanding feature, that is network connection, allowing people to have potential access to an infinite amount of data in potentially dangerous situations. These data are now accessible in most cases from external devices such as smartphones, tablets, satellite navigation systems and multimedia data storages.

On a sample of 118 people aged between 18 and 65 years, 80% of those interviewed said they constantly use these objects; out of these, 65% of them admitted that sometimes, while driving, these devices distract or confuse them.

Following this first analysis, researchers tried to model users' behaviour through the use of questionnaires and the Personas method.

Four Personas were identified and 8 relevant factors were taken into consideration: technological attitude, driving style, mileage, destinations, cognitive load, approach to mobility, level of education and lifestyle.

The questionnaires and the Personas method proved to be excellent design tools, despite being very different from a real ethnographic observation in the field. In this way, it was possible to draw a frame of users' perceptions and to obtain a number of indications about possible uses and wishes.

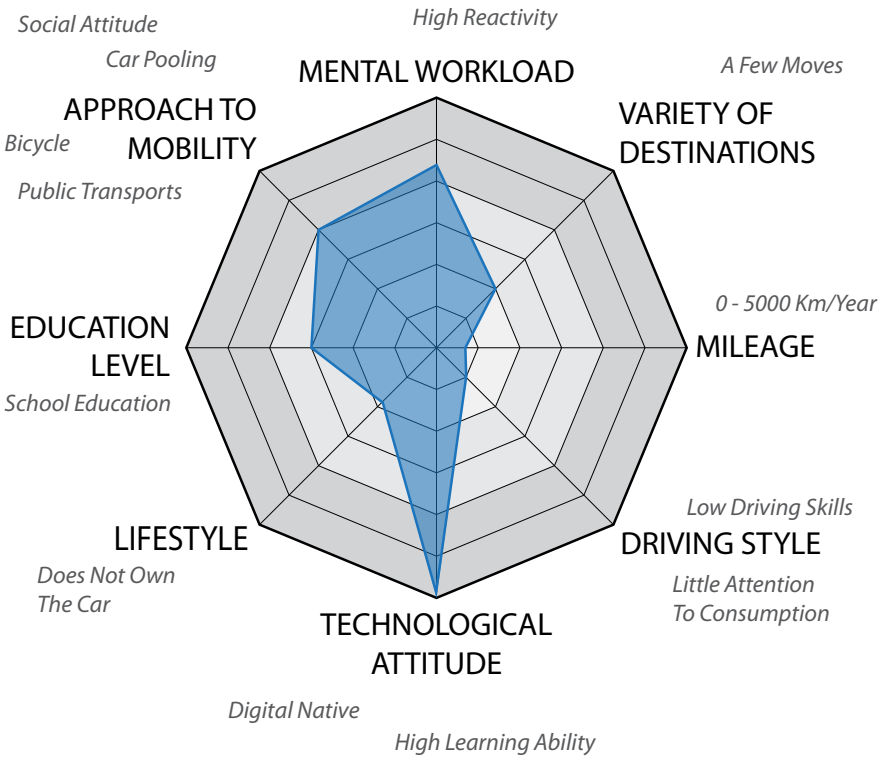


Fig. 6. Eight relevant factors of Personas

4 Project

The designed interface consists of two components: a HUD showing quantitative information that can be read and understood with quick eye movements, and a HDD that shows qualitative information with more detailed data.

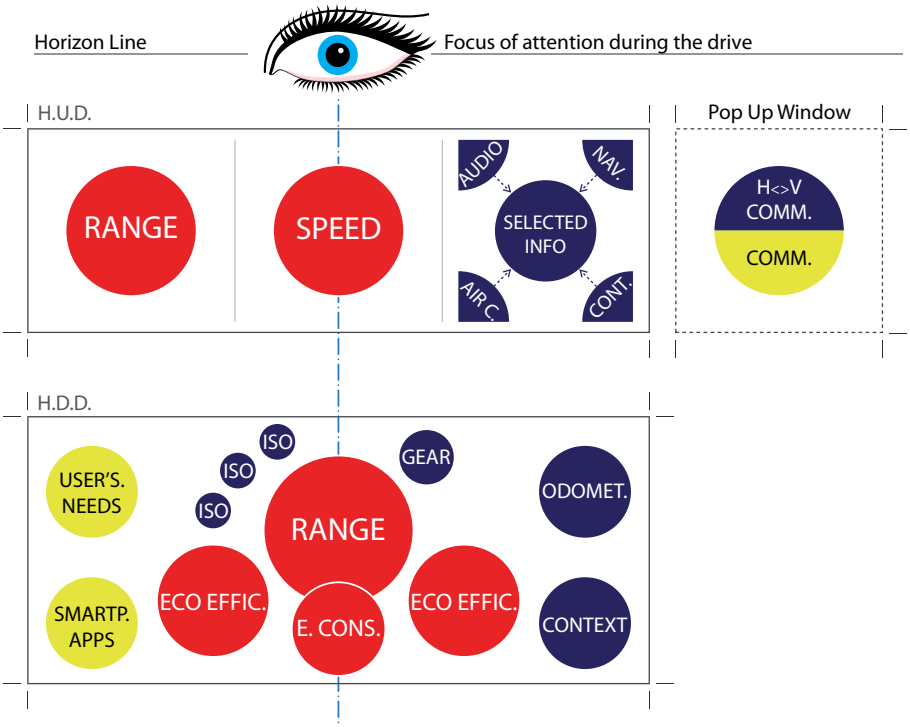


Fig. 7. The new designed interface

The choice to divide the display into two parts derives from two considerations: first of all, the driver's focus of attention, despite being in constant motion to scan the surrounding environment, as necessary, remains on the road during most of the driving; secondly, the results of a test confirm that the driver's reaction time to potentially dangerous situations is one second shorter using HUD than using HDD (1,323" with HUD and 1,327" with HDD) [R13].

The HUD is divided into three parts, plus an optional fourth one that the driver can retrieve. The three parts that are always present show the basic data for a safe drive of the vehicle: speed at the centre, state of charge on the left, and audio, air-conditioner and navigation data on the right. The fourth portion of the HUD is the most interactive one and it contains information about possible emergencies, driving suggestions from the system, synchronization with nomadic devices (like the agenda) and communications with the external environment (like phone calls but also e. g. points of interest).

Today, secondary information is often arranged in the central area of the dashboard. In this system, it is concentrated in an accessible area, open to immediate vision, eliminating the need for the driver to rotate the head. In the HDD, instead, pieces of scalable and customizable information are arranged around basic information: the state of charge, when the car is moving, or eco-indicators, when the vehicle is switched off.

Other information can be added according to user's demands and characteristics, leaving of course the appropriate space to indicators required by law. Visualization is based on an interaction that is completely different from the currently available one: in fact, it should be totally blind and based on consolidated gestures. This type of interaction is very simple and due to its nature it clearly limits excessive branching in menus, thereby preventing the proliferation of unusable menus and submenus. As already said, information has been divided into three categories, however it is necessary to allow a certain degree of customization according to emerging differences in Personas features, self-learning ability of the system and present needs, still preserving priority and security. A person who has just obtained a driving license, for example, will be willing to receive more suggestions, will need to visualize potentially dangerous parameters in a clearer way and will want to connect his or her technological devices to the system, which will filter information according to the driving style.

The system is programmed to react to three types of inputs: the voluntary ones, the unconscious ones and synchronization inputs. Voluntary inputs are specific requests of the user who needs to obtain some specific information in a precise moment or to change the display in order to suit his or her vision needs. HUD and HDD settings can be calibrated to change focus, size, brightness, contrast and saturation parameters, in order to obtain a dashboard that is not only usable but also accessible.

Unconscious inputs are defined like those parameters that can be detected by the system, for example through sensors and data analysis. Physiological parameters of the user can be monitored through sensors on the seat belt, for example, analyzing the heartbeat to perceive stress, or through cameras, analyzing the dilation of the pupil and the number of blinks. The on-board system instead can analyze the driving style and judge it as too nervous and careless of fuel consumption, for example, if it detects sudden cornering, acceleration and braking, or if it perceives excessive pressure in commands activation .

Synchronization inputs represent the chance to connect the car system to the agenda and to possible requests not only by the user, but also by his or her network of contacts.

The system is designed to learn from user's behaviours, and clearly the more it will be used, the more it will become reliable and pleasant to use. At the same time, it does not want to be an annoying assistant and the basic state can be restored at all times.

5 Open issues and future works

Although the whole project has not been evaluated by a remarkable number of users yet, the adopted methodology is based on an iterative process which is characteristic of human centred design. Future works will deal with the new display mode, its acceptance and the evaluation of usability and accessibility, even in critical driving phases.

References

1. Ensdley, M. R. , Garland, D.J.: Situation Awareness, Analysis and Measurement, Mahawah Lawrence Erlbaum Associates, SA Technologies (2000)
2. Norman, D.A.: Emotional design: why we love (or hate) everyday things, Basic Books, New York (2004)
3. Di Salvo, A., Gaiardo, A., Ermacora, G.: MOBIOT: Sustainable social mobility in the internet of things. In MOTSP 2013 5th International Scientific Conference, Management of Technology Step to Sustainable Production, Novi Vinodolski, Croatia, 29-31 May (2013)
4. Bistagnino, L. Systemic Design: Designing the productive and environmental sustainability. Bra (Cuneo): Slow Food Editore (2011)
5. Preece, J. Interaction design, beyond human-computer interaction. New York: John Wiley & Sons (2002)
6. Germak C. (ed.). Man at the Centre of the Project: Design for a New Humanism. Turin: Umberto Allemandi & Co (2008)
7. http://awards.ixda.org
8. http://www.bmw.com
9. http://www.audi.com
10. http://www.mercedes.com
11. Neureiter, K., Meschtscherjakov, A., Wilfinger, D., Tscheligi, M.: Investigating the Usage of Multifunctional Rotary Knobs in the Center Stack with a Contextual Inquiry, Adjunct Proceedings of the 3rd International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Salzburg, 127-142 (2011).
12. Tönnis, M.; Broy, V. & Klinker, G. 2006. A Survey of Challenges Related to the Design of 3D User Interfaces for Car Drivers. 3DUI '06: Proceedings of the 3D User Interfaces, IEEE Computer Society, 127-142.
13. Liu, Y.-C.: Comparison of Head up display vs head down display: driving performances of commercial vehicle in Taiwan, International journal of Human-Computer Studies (2004)

Shaping Organizational Complexity

Azzurra Pini and Paolo Ciuccarelli

DensityDesign Research Lab, Politecnico di Milano,
Design Department, Milan, Italy

azzurra.pini@polimi.it, paolo.ciuccarelli@polimi.it

Abstract. The increasing complexity we are experiencing at a societal level is undoubtedly and inevitably affecting also organizational systems, with emerging issues related to their specific function. The vast amount of data available thanks to the current practices and technologies became a ubiquitous factor and is giving rise to a whole range of complementary questions related to the way we access information, that in complex environments always occur in a selective way. Both the scientific community and organizations are therefore drawing their attention towards the importance of visualization and the benefit of design tools for organizational systems. Nevertheless, theoretical research is rarely applied and remains basically confined within the disciplines where it has been conceived. Similarly, several techniques and tools have been developed in order to deal with organizational complexity and support decision making through models that attempt to highlight the structure of the system. Current tools though, result in technical representations, applying very rigid and standardized models that rarely convey the complexity and dynamism of the system, thus providing a sole and strained access point to multidimensional information.

Considering organizational complexity as the result of an observation of information and social complexity we also highlight the distinction between the formal and informal aspects of organizations. In this paper we present two different research projects, with the aim of visually exploring both aspects of organizational complexity within large business companies. The first project (SICHT) examines the human and social aspect of organizational life, which is strictly related to social roles and the network of informal relationships between individuals. The method most commonly used to understand staff social networks is Social Network Analysis, a technique that employs software tools that lead to graphical analysis summaries in the form of static graphs. The project presents the design of a visual interface, suggesting the utility of a diverse way of visualizing the results of such an analysis, through an interactive and more flexible model, with different display patterns to reveal peculiar aspects of the data. The second project (NOTON) focuses on information complexity, which is directly related to the problem of information accessibility and materializes in particular into formal organizational processes and workflows. NOTON concerns the design of a visual interface for the exploration of innovation processes. By exploiting innovation specific properties, analyzed through its objectives, context, time dimension and uncertainty level, the project outlines some possible directions to be explored for the design of process visualization tools.

The proposed projects represent an opportunity to highlight the importance of design principles and techniques for the visualization of organizational constructs, both at a social and informational level. By claiming a diverse approach to the design of tools for organizations we focus on the complex, evolutive and ambiguous nature of organizational data. In particular, this contribution aims at emphasizing the need of a more prominent role for the design disciplines such as communication design, with specific reference to information visualization and interface design, to function as a bridge between different research fields.

Keywords: information visualization / organizations / interface design

1 Introduction

Besides the implications of complexity theory have been studied for quite a long time, the application to organizational life came just recently, together with the awareness of the uncertainties associated to activities and communication. Emerging difficulties in decision making have been attributed by organizational studies to a bounded rationality principle, which introduces the concepts of risk and information cost and considers the empirical nature of the decision process. [9]

The problem of information accessibility is recently becoming particularly relevant as the possibility to automatically collect and store large quantities of data and information of various kinds is by now a current practice in organizations. [16] However, these huge repositories remain largely unused and inaccessible to decision makers and stakeholders in general. Although the problem has been explored by several disciplines, claiming for methods and tools to support decision making, most of the current research does not take in sufficient consideration the conceptual and operative nature of the design discipline, with specific reference to diagramming techniques. In our vision, organizational systems and diagrams share the same original purpose, that is structuring and managing complexity. Diagrammatic representations embed the capability of shaping and communicating complexity by abstracting form in time and space, overcoming language constraints of different disciplines by giving a visual shape to a problem space. The interaction with a visual interface represent an additional mean of establishing a connection between the different aspects of a complex system such as organizations, by stimulating a mechanism of knowledge amplification and consequently a support for decision making. [15]

Our considerations start from the assumption of organizational complexity characterizing knowledge production and knowledge sharing, and consider the large number of elements and their rich and nonlinear interactions over time. Multidimensionality of elements and the overall uncertainty that characterize data and information associated to them is therefore an intrinsic characteristic of all organizational constructs. Social and informational complexity represent two aspects of the same phenomena that is usually studied by different disciplines autonomously. These conditions reflect the distinction between the two sides

that compose organizational systems, which are respectively informal organization and formal organization. [8] While the formal side comprehend the official and visible structures of the organization, such as roles and processes, the informal one is related to the human component and the communication that exist between members. Both aspects contain numerous sources of ambiguity and complexity, which have been explored through two different projects that allowed to develop some considerations on representational issues for complex systems.

2 Visualizing the Informal Organization

While dealing with the visualization of the organizational structure, one of the main sources of complexity is the social component, which reflects in the form of the informal organization. The intricate network of relationships existing among participants is not usually created on the basis of official criteria such as roles and hierarchies, but of subjective principles that establish a communication preference, related to concepts such as affinity and trust. The formal side or the organization, while defining roles and procedures, create the institutional foundation of all activities and communications, but numerous studies demonstrate that 79% of the total workflow is actually performed through the connections established by the informal organization. [14]

The representation of social connections in groups has been studied in social sciences starting from Moreno's sociometry in 1953 [12] and the lattice structure is considered by numerous disciplines as a preferred solution for the representation of complex systems in general. [4] The current Social Network Analysis (SNA) [5] is the discipline that operates between sociology and statistics that studies the structure of relational data in social contexts, by considering the characteristics of the single units as emergent from the relational structure. Mapping the interaction between individuals helps the understanding of the system behavior and in Organizational Network Analysis (ONA) is often used as an indication to produce modifications in the formal structure in order to match organizational goals [14]. The network analysis performed in organizational contexts generally implies administering a survey to employees, which results are organized and represented with specific software, such as Ucinet and Netdraw. The resulting visualizations though are generally presented to stakeholders in the form of printed reports, thus eliminating the fundamental concept of dynamism and constant transformation of the social network. Moreover, the representation of the relational structure doesn't go beyond graphs, even if they represent just one possible solution among several different visual models [3]. Another important aspect concerning the visualization of social networks is the multidimensionality of data, which reflects the complex nature of the phenomenon. The possibility to explore the social structure from different points of view and with different levels of detail allows both a better understanding of the phenomena and knowledge creation. Several studies also demonstrate that multidimensional stimuli produce better results in cognitive performance. [11]

2.1 A Case Study: SICHT

The SICHT project has been developed as a conceptual design in collaboration with an Human Resources consultancy in order to analyze the peculiarities of an organizational social network and to explore alternative visualization techniques and interaction patterns capable of providing a dynamic and meaningful way to access the results of ONA. The analysis highlighted the necessity of a tool with specific characteristics:

- *context specificity*: most of the tools commonly used for the visualization of organizational networks are generic and are directed to analyst instead of stakeholders and decision makers;
- *interactivity*: due to the scarce usability of the current tools and their overcomplicated interfaces, HR consultancies generate static reports that obviously can't be manipulated and explored by final subjects;
- *time*: taking into account the unstable and spontaneous nature of relationships that bud between people, the structure of the social network needs to be analyzed over time, both to program and to observe the effects of the informal events that fill organizational life;
- *perspective*: considering the organization, and therefore its network of relationships, as a complex entity, its social structure needs to be considered from different points of view, providing diverse visualizations to highlight different aspect of the phenomena.

The interface. After a first step that concerned the analysis and the elaboration of the dataset provided by the consultancy in the form of square matrices [17] (the typical format for network data), the interface has been structured in three views, which reflect the main perspectives of organizational social networks according to the interviews conducted with HR managers of different companies.

The main interface allows to explore the different nested levels of the social network, that articulate from a superficial "know network" to a deeper "trust and problem solving network". The possibility to filter the different attributes is always present in all views, as the indication of the network metrics, such as the tie strength or the path length (the number of intermediaries between two given nodes). The evolution of the network structure can be explored along the time dimension in the bar at the bottom, where different organizational events help to keep track of the possible situations that could modify the social structure.

The first *Network* view reflects the typical graph pattern used by organizational network analysis. The nodes have been represented by combining the three main qualities of the resources (role, department and base) into a single glyph, in order to keep the overall view of all node attributes. The filtering panel on the left allows to select the attributes that the user wants to compare, simplifying the visualization by eliminating one or two of the three node layers, thus allowing a better comparison.

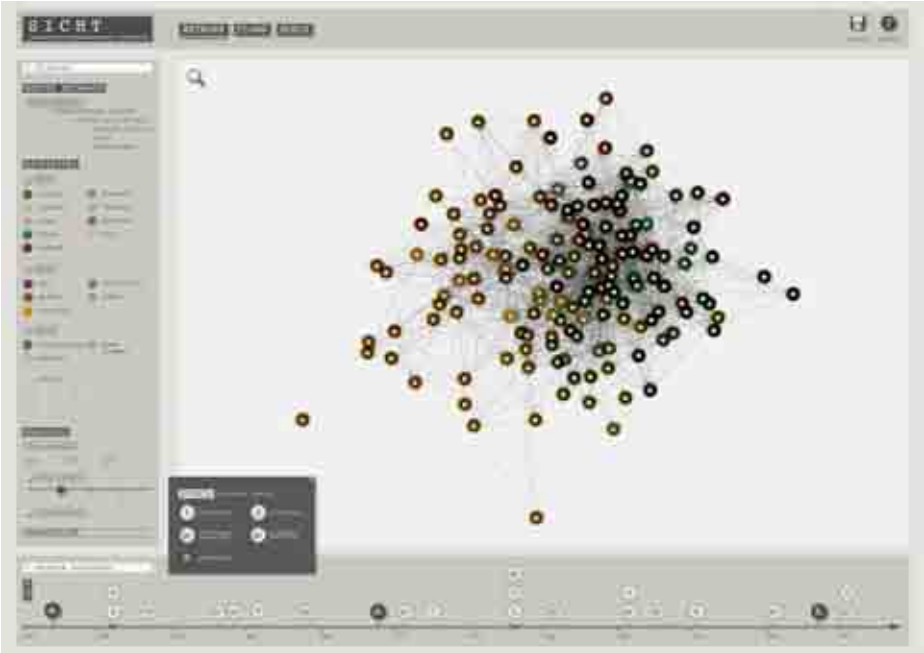


Fig. 1. SICHT: The Network View

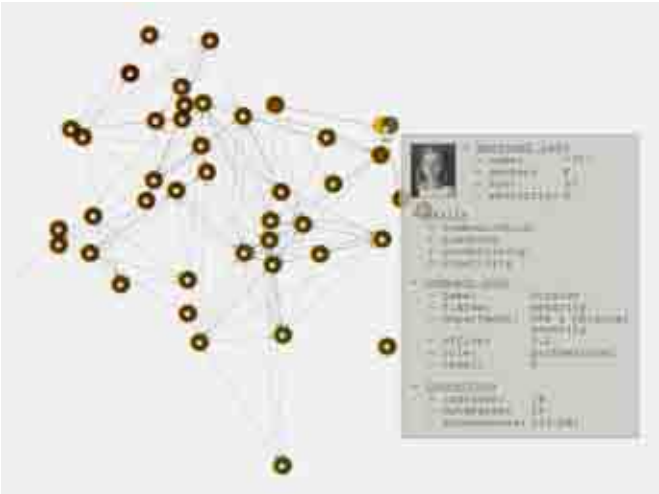


Fig. 2. SICHT: A node expansion

The *Flows* view highlights the departmental network, through a sunburst visualization model [18]. Data have been hierarchically aggregated according to the base, department and role of each employee. Each ring represents an attribute of

the personnel and the order of aggregation can be set through the filtering pane from the outside to the inside. By selecting single blocks the user can then discover the relationships between specific areas of the organization.

Fig. 3. SICHT: The Flows View

Fig. 4. SICHT: an example of flow selection

The third view *Space* concentrates on a geographical dimension that generally is not considered by the tools for social network analysis and represents a way to compare the nature of relationships with the actual distance that separates people. The network is visualized through a quantum treemap [2] for each base, subdivided into offices and coloured by department. This visual pattern allows to keep the single units within the aggregated representation, providing also an indication for space planning.

Fig. 5. SICHT: The Space View

3 Visualizing the formal organization

The representation of formal organization can be assimilated to the representation of processes, a practice that has been a subject of research for a long time in different disciplines, mostly related to engineering and computer science, developing several methods, such as flowcharts, state transition diagrams and Petri nets.

While business processes have been studied for a long time both in organizational sciences and computer science, innovation processes don't benefit of the same attention, they are studied in different fields and don't have the use of adequate ad hoc software tools to represent and manage activities. [19] Current tools for process visualization, while applying to business processes, mostly concentrate on performance and consequently on process enactment, following a

quantitative approach and looking for data accuracy. Simultaneously the development of notation standards, such as UML, have had undoubtedly a functional role, to enhance interoperability and to simplify entities definition, but results too rigid and constrictive for knowledge-intensive processes. Processes related to innovation activities, as oriented to the exploration and dealing with uncertainty, feel in fact the need of a more loose set diagramming rules, with the aim of showing process inefficiencies and addressing process redesign.

By examining the peculiar properties of innovation processes, this research relate them to the design of a visual interface for the exploration of innovation processes.

3.1 A Case Study: NOTON

NOTON is a research project that the research lab has been carrying on with the innovation department of a manufacturing company. The general aim of the project is the creation of a shared representation of the innovation process among the different stakeholders, through the design of a visual interface to explore the process itself.

Dealing with the company datasets and confronting with the innovation department, we were able to draw some considerations on different dimensions of innovation processes, which characterize specifically the examined project:

- *objective*: being a multi-actor system, innovation processes need to be explored through multiple perspectives, with a different level of analysis and different goals;
- *context*: innovation processes are more based on projects then task execution. Projects articulate in phases, delimited by milestones, which are likely to be to be skipped or abandoned at any stage;
- *time*: innovation projects live a different time dimensions, as they can last for years or for a few weeks, depending on different conditions (market push, technical feasibility, etc.). They can also experience iterations and contain feedback loops;
- *uncertainty*: the nature of information is generally qualitative and many properties are ambiguous, not defined through a single or a closed set of indicators, but described along different dimensions [10].

The project we are presenting here is to be considered a preliminary mock up which is not intended to respond to all issues and challenges presented, but serves an exemplifying purpose, to confront how some of the concepts highlighted can be observed and tackled by means of visual design.

The Interface. Considering the preliminary stage of the project, a first activity has focused on the abstraction of the process structure, in order to highlight the critical aspects to be discussed with the innovation department. The characteristics of the specific context reflected in the interface design by concentrating more on the reproduction of the overall view than on the analytical and operative function. We decided therefore to adopt a system of alternative multiple views, to be distinguished from multiple coordinated visualizations [1], in order to enable the exploration of processes through different perspectives, representing different specific communication goals.

As the original representation of the process used by the company was a static funnel chart containing a selection of projects for each period considered, the first concern of the interface design was the dynamic representation of the time dimension. A specific need was the possibility to explore projects over time, both in analytic and aggregated way, in order to see the trend of projects along certain periods, and detect possible spaces to intervene with strategic actions. In Business Process Modeling Notation (BPMN) [13], the discipline that studies the graphical representation of process models, processes are generally represented as a sequence of tasks, and their specific properties are defined within the notation itself. NOTON instead is working on the intersection of two different temporal dimensions: the analytic one, which follows the regular course of time, and the structural one, which concerns the passage from a phase to the next one. Observing the process progression simultaneously through these two different dimensions, allows to create a visual comparison and to assess the compliance of the two trends.

Within the *Timeline* view (Fig. 6, 7), the analytic visualization at the center displays the projects through a line chart. On the x-axis we plotted the time dimension, to display the singular dates for each milestone, which are shown on the y-axis as an ordinal variable. Visual perception experiments discovered in fact that spatial position as in a scatter plot, bar chart or line chart, facilitate a better understanding of data, with respect to other visual variables such as angle, length, area, volume, and color saturation. [6] Every dot in the line chart defines the date of transition from a phase to the other, which is described by a milestone (M0, M1, etc.). The data however show a certain complexity, registering up to two different dates for each milestone, depending on the accounting accuracy for each project. The first date is described as the “expected date” and represented as a smaller light grey square, while the “real date” is displayed as a bigger and darker square. As a first indicator of uncertainty, the use of a line chart plotted on two axes allows to draw attention towards the compliance of the project path with respect to the expected one. The shifts from a regular straight line going from bottom left to top right to a broken line, can indicate in fact a return to a previous step, possible feedback loops or some kind of data inaccuracy.

Fig. 6. NOTON: Timeline view with filtering

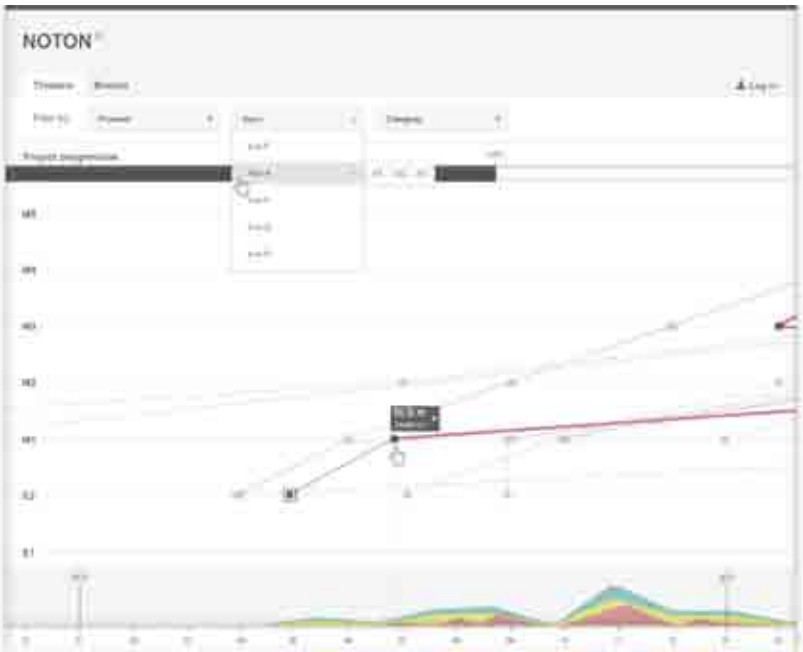


Fig. 7. NOTON: Timeline view with zooming and highlight of risk level

The *Stream* view (Fig. 8) provides the big picture of the process, by aggregating projects by type and by level of risk in a stream graph. The milestones here are switched and displayed in the x-axis in order to establish a connection with the time meter at the bottom. In this case the previous timeline has been changed in favor of a more abstract one, to allow the measurement of the average duration of each phase in terms of weeks or months, selecting the period of interest from the top menu.

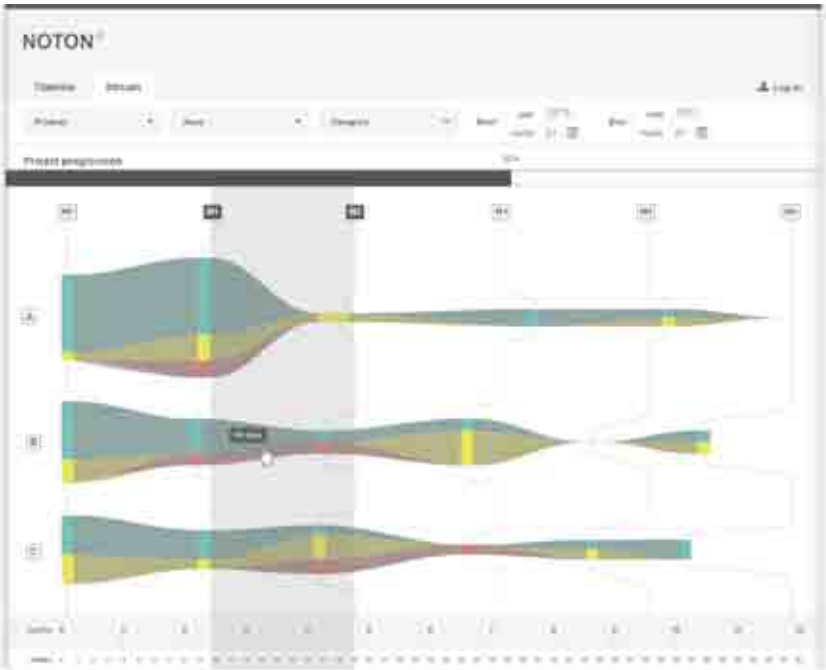


Fig. 8. NOTON: the Stream view

Conclusions

Our contribution tries to set a new possible and profitable area of applicability of communication design within organizations, with specific reference to the design of visual interfaces.

The complexity of organizational systems has been observed separately from two different points of view, the social and the processual one, in order to analyze the peculiarities and the analogies of the two aspects to be expressed in the design of visual interfaces as decision support systems.

Constitutionally design operates with the aim of motivating the emergence of new perspectives and different modes of observation, through a nonlinear approach. Thus we claim that researching on the modality of visual representation

through interactive interfaces would be particularly valuable both as a support for decision making and as a comprehensive communication tool.

The possibilities for a tool capable of visually represent the structure and the evolution of complex phenomena such as informal relationships and innovation processes would be important not just as a support for current decisions, but also as an indication for organizational redesign and as a representation and communication tool for organizations.

References

1. Baldonado, M.Q.W., Woodruff, A., Kuchinsky, A.: Guidelines for using multiple views in information visualization. In Proceedings of the working conference on Advanced visual interfaces. ACM (2000)
2. Bederson, B. B., Shneiderman, B., & Wattenberg, M.: Ordered and quantum treemaps: Making effective use of 2D space to display hierarchies. AcM Transactions on Graphics (TOG), 21(4), 833-854 (2002)
3. Bertin, J., & Barbut, M.: Sémiologie graphique: les diagrammes, les réseaux, les cartes. Paris: Mouton (1967)
4. Capra, F.: The web of life: A new scientific understanding of living systems. (1996)
5. Freeman, L. C.: The development of social network analysis. Vancouver: Empirical Press, (2004)
6. Heer, J., Bostock, M., & Ogievetsky, V.: A tour through the visualization zoo. Communications of the ACM, 53(6), 59-67 (2010)
7. Kilduff, M., & Tsai, W.: Social networks and organizations. Sage (2003)
8. Luhmann, N.: Organisation und entscheidung, Opladen: Westdeutscher Verlag (1978)
9. March, J. G., & Simon, H. A.: Organizations (1958)
10. McManus, H., Hastings, D.: A framework for understanding uncertainty and its mitigation and exploitation in complex systems. IEEE Engineering Management Review, 34(3), 81-94 (2006)
11. Miller, G. A.: The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological review, 63(2), 81 (1956)
12. Moreno, J. L.: Who shall survive? Foundations of sociometry, group psychotherapy and socio-drama (1953)
13. OMG Business Process Model and Notation, http://www.bpmn.org/
14. Oriani, G.: La forza delle reti di relazioni informali nelle organizzazioni. L'Organizational Network Analysis, Milano: Franco Angeli (2008)
15. Pike, W. A., Stasko, J., Chang, R., O'Connell, T. A.: The science of interaction. Information Visualization, 8(4), 263-274 (2009)
16. Reichert, M.: Visualizing Large Business Process Models: Challenges, Techniques, Applications. TAProViz12 pre-proceedings, Tallinn (2012)
17. Scott, J.: Social Network Analysis. A Handbook. London: Sage Publications Ldt. (1991)
18. Stasko, J., & Zhang, E.: Focus+ context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. In Information Visualization, 2000. InfoVis 2000. IEEE Symposium on (pp. 57-65). IEEE (2000)
19. Van Der Aalst, W. M., Ter Hofstede, A. H., Weske, M.: Business process management: A survey. In Business process management (pp. 1-12). Springer Berlin Heidelberg (2003)

Small Data Visualization Toolkit

Daniele Savasta

Iuav University of Venice, PhD student in Design Sciences,
Venice, Italy

daniele.savasta@gmail.com

Abstract. In this study I describe an open-source toolkit for the visualization of small data. Three main areas are discussed in the paper: artifacts, data and people. Today, communication is supported by numerous interactive artifacts. The omnipresence of interactive artifacts increased the production and the storage of data. The huge collections of data produced, mainly online, are defined "Big Data". Opposed to this, a more recent concept has appeared: the "Small Data". Small Data are roughly defined as "everything that can be processed in Excel". In order to reach at a more operative definition it is appropriate to define the characteristics of the Small Data as distributed, detailed, decelerated and democratic. These characteristics make Small Data look more understandable, manageable and in the end useful for common people. It is a common interest to understand which data can be useful for people, which are recorded, and who is using them and for what purpose. Designers should work more on enabling platform to help people transforming data into knowledge. A toolkit proposal could be useful to highlight some priorities that designers should consider for a better involvement of the people.

Keywords: small data / social interaction / generative design / design toolkit

1 Introduction

Human being started to record information around 8000 years ago through writing and painting. By the time, the message's carrier changed several time from stones to walls, to clouds and bones, to architecture and paper. Today we are back to cloud. The aim of the Internet at his birth was the exchange of information. A feature of the so-called information age is the capacity of data storage and the easier production due to diffused presence of computers. Since Arpanet to today, in less than fifty years, the net spread all over the world and with new artifacts, sensors and services the amount of data produced augmented to 2.5 quintillion bytes daily [1].

Today, numerous interactive artifacts support communication. Manufacturers embed computers and sensors in products that were not conceived as computational devices before. The more classic examples are fridges and washers, but also cars are "mobile observatories" [2] as smartphone, tablet, eBook readers

and so on. The number of products that embed some computer is continually increasing with the aim of improving, at least nominally, people experience.

We ended up having a symbiotic life together with these devices [3], but I believe that this continuous attachment and this dependency are unbalanced and clumsy. Devices and artifacts surround people's life recording it and 'helping' them to keep communication with others humans. Now, we should ask ourselves if these computers are help or obstacles. Artifacts change people's behavior and today it is not uncommon, and often not perceive as unpleasant, using a smartphone in any occasion from metro to meeting, from classrooms to the dining table. They are always on forcing people to be up to them. Their capacity of catching attention and absorbing time are immoderate. Sometimes the continuous distraction becomes an insult to the people physically around [4].

Nonetheless, people use mobile devices to share their life, show emotions, spread ideas, sometimes still for call and speak with someone and keep contacts with people that are not next to them. These devices compress space and time around us, so it is much easier to catch up with people on the other side of the world than with the ones right next to us [5]. It is a fixed binocular on eyes; distant friends look closer, but who is next to us disappear in blur. People get so accustomed to these prostheses that we feel their absence when we lost them or when they just go out of charge. We suddenly become alone. Disconnected from the world we cannot share information anymore.

This can be correct only for the conscious information, but there are a lot of unconscious ones that we still leave as tracks of living, from the use of credit card to the last cell that recorded the phone. Many are the data that we produce through devices without even being conscious of them. Sometimes these data can be used immediately or recorded. A use for stored data can be found later, from the company or for other companies interested and allowed to access the data [2]. "Big Data" is the definition used for the large amount of data owned by large companies.

Next paragraphs will discuss what are Big Data, and consequently a proposal for a Small Data's definition. The definition of Small Data will make emerge the problem of Small Data visualization. In the end, a practice example will show a toolkit that try to make Small Data understandable by people which is the aim of this study.

2 Big Data

The concept of Big Data comes in our days as natural consequences of the increasing amount of data produced thanks to the recent digital technologies. Big Data can be traced back to fifties, but only recently Big Data received great interest, mainly because of the economical results and more obvious examples from Google Flu Trends[1] to Ford Silicon Valley Lab[2] just to name two of them.

[1] Using the data collected from the research on his search engine Google developed an algorithm to analyze the diffusion of influenza.

Big Data it is "a meme and a marketing term, for sure, but also shorthand for advancing trends in technology that open the door to a new approach to understanding the world and making decisions" and for this reason it is relevant to analyze and try to understand it better [6]. Big Data can be defined by three main characteristics employed by Laney: volume, velocity and variety [7].

Volume. Volume is the first and maybe more significant mark; Big Data are big, if 2.5 quintillion bytes is not already impressing, it is the same of saying 576 billion times the bible, 531 million DVD, 2 million of commercial 1Tb hard drive and so on. This is the dimension of Big, and obviously this means that not everyone can store them. Only large companies can afford these specifications. The sources of this large amount of data are the movements of GPS, credit card transactions, social network updates, photos, videos, and metadata recorded in forms; but also click on web-pages, the time spent on a page, the appreciated books and the music listened or produced. Friendship is been digitalized. Shopping online and fidelity cards record the people's preferences. Everything is stored somewhere.

Velocity. People cannot manage these large amounts of data without computers, and the computational power requested is huge as the storage capacity. Velocity is the second character. The rapidity of digital life, but most of all the quickness of market, asks for instant answers. Data analysis needs to follow the same speed, also because many of the mined data involve real-time variables and a real-time response can be required.

Variety. All these collections can be also unpredictably various. Data mined from social network can contain relevant professional photo shoot as the worst auto shoot, from the author's cite to irrelevant sentences, from musicians to unlistenable track and everything is in between and all of them together. Data can be structured or not they can be in a different form and contain different modalities. This variety is the third and last of the main characters that define the "Big Data".

Other authors try to add V in order to update the Big Data definition[3]. Veracity, contested by Laney in a comment of IBM's infographics [8], is an interesting mark that shows not a value but a birth defect that comes with volume. Veracity is the uncertainty of data. On a widespread data mining, there is no space for selection; all the data are gathered and collected, mixed and stored adding quantity and inevitably subtracting quality from database.

---

[2] Ford has a Big Data analyzing center to research on improvement or problem solving for his cars thanks to the data received from the multiple sensors installed inside cars.
[3] Other than Veracity, online it is possible to find Viability and Value http://www.wired.com/insights/2013/05/the-missing-vs-in-big-data-viability-and-value/ and http://www.pros.com/big-vs-big-data/, a discussion on the topic is proposed by Seth Frimes in the article Big Data: Avoid 'Wanna V' Confusion on http://www.informationweek.com/big-data/commentary/big-data-analytics/big-data-avoid-wanna-v-confusion/240159597

Big Data can be a dream for marketing where large amounts of number seems to match with huge amounts of money, but is not easy to work on these data. Volume has his pros and his cons: only "Big companies" own resources to store and manage Big Data. Recently, starting from this observation, and maybe trying to surf the same wave of Big Data, the idea of "Small Data" emerged [9][10]. Under the following title, we will see some examples, and I will try to reach at a definition for this new concept.

3 Small Data

Rufus Pollock, economist and co-founder of the Open Knowledge Foundation Labs, define Small Data as "everything that can be processed in Excel" [11]. This sentence underlines the opposition in terms of computational power between Small and Big Data. Clearly the aim is to propose something that everyone can process on his own computer without any special equipment.

Small Data is also an expression used to define what was before the Big Data [5], but that would be a restricting definition. Was it a small data world because people were not measuring it? Was a small data world because people did not have the tools to elaborate large amount of data?

It looks more reasonable to talk about Big Data of the past, maybe through the analysis of archeologists and anthropologists. Big and small data do not seem divided by the time. In fact, Small Data's use in the last months shows some substantial marks in opposition to Big Data. The computer scientist Deborah Estrin shows how individual digital traces help for medical care [12]. All the data recorded can shows not only medical state but also behaviors, helping doctors to understand how a treatment is affecting us or what can be the cause of an illness. She claim that all the data generated by mobile technologies and recorded by search engines, social networks, internet providers, mobile carriers should be given back to the person that generates them. Small Data is a piece of Big Data that refer to ourselves.

Pollock define Small Data as "the amount of data you can conveniently store and process on a single machine, and in particular, a high-end laptop or server" he proceeds: "What's interesting (and new) right now is the democratization of data and the associated possibility of large-scale distributed community of data wranglers working collaboratively. What matters here then is, crudely, the amount of data that an average data geek can handle on their own machine, their own laptop" [11].

I believe that after this analysis a definition of Small Data can be expressed through these marks: distributed, detailed, decelerated and democratic. These characteristics make Small Data look more understandable, manageable and in the end useful for common people.

Distributed. Small Data are by chunks of relatively small amount of data that a single person or small groups can store and manage. The distribution of data is conceived as he opposite of the volume the concentration of different data from

different sources in databases owned by single large companies. I can decide to share my own data with my doctors, lawyers or friends and coworkers for functional reasons or for recreation, creating my networks.

Detailed. Small Data are specific to a context. This selection at the beginning avoid useless or too generic data. The data collected in Small Data are coherent and solid. Their definition and size allow an easier check of anomalies and the resolution of them.

Decelerated. The quickness requested in Big Data can be softened in favor of a better certainty. The calculation power requested to process Small Data are by definition moderate and fulfilled by common commercial machine. This mark makes Small Data decelerated.

Democratic. Democracy is a controversial term, but here we mean that the Small Data are thought to be owned by a person who generates them, not by the company that supplies the service or product in use. The democratic value of Small Data is strictly related to the idea of sharing and collaborating, giving to the people the freedom to open, manage, store, edit, visualize and share their own data.

Designers should start to think to help people in this change. Small Data are definitely smaller than Big Data, but it is still impossible to understand them without a proper visualization. In the next paragraph, I will propose a toolkit developed starting from these observations.

3 A toolkit proposal

Information for Social Emancipation. Collecting and producing data does not mean create either knowledge or information. Data is information translated into a form understandable by computer. Knowledge is the understanding of information for human beings. So information can be seen as a communication level between computer and human beings. Designers have to shapes data in a convenient way for humans, through information visualization. As Bonsiepe pointed out [13], this interest in information visualization can be just a growing trend, it can be senseless consumption of information for instant satisfaction, but it can also be a growth in consciousness for the importance of information. Sharing information can be a way for emancipation that needs a coherent, understandable and pleasurable visualization.

Towards the definition of the toolkit. People are trying to obtain the data that they should own already, but a real understanding of the information will be missing without a proper visualization. In this study, I will try to define a toolkit for people to understand data through data visualization. The analysis of several different toolkits for data visualization, already available online, has raised a

main issue: they are developed for programmers or skilled designers, not for common people. The selected toolkit for the comparison were: jscharts 3.0, Harry Plotter v0.9, Toxiclibs, gRaphaël's, morris.js, MilkChart, jQuery Sparklines, JavaScript InfoVis Toolkit, D3.js, Highcharts 3.0, Flotr2, Google charts,. Most of them are released under the MIT license so they are open source, but the requested skills in order to use them are far beyond the medium level skills of a common person. Specific skills in programming, mainly in javascript and html, are requested. Some of them like D3.js, Google Chart and Highcharts 3.0 offer a wide choice of graph and diagrams; other are more specific like the Tufte's ispired jQuery Sparklines. The possibilities of creating interactive and/or animated chart are common but unlike some of them claim "good-looking charts shouldn't be difficult", actually it is. The emerging needs of data visualization and these observations defined the aim of the project.

The purpose of Small Data Visualization Toolkit is to enable a low skilled user to transform data to visualization with simple steps: data loading, purpose's selection, visualization of results, customization and saving. The division of the screen into three portions allows the person to see all the procedure at once. An enlightenment of the title will suggest the focus. A dark grey pattern will overlay the other portions. Proceedings trough the phases will turn on the next portion and put in a stand-by the previous one. Because it is useful to move back and forward during the editing, a light white overlay will weaken but not hide the edited portions. The effort is to keep the screens as simple as possible, to make people feel to be in the middle of an easy process.

Data Loading. In the first portion of the screen starting from left, there will be the data loading process. Data can be loaded on the toolkit platform by simple drag and drop the file on the portion of the screen designated. Many software and services offer the same procedure making it widely known by common people. Otherwise, a click on the upload data area will open a dialogue screen that allows choosing the file. The toolkit accepts the most common formats like the comma-separated file (csv), or Microsoft Excel (xls). A direct connection is allowed with a Google Drive Spreadsheet. When data are loaded, the first portion of the screen will show the content of the file in a spreadsheet like mode. A small bar on the bottom will appear to allow going back and changing the data selected.

Purpose's selection. The portion in the middle of the screen will be showing some questions relative to the purpose of analysis. Supposing that the person does not have any skill I thought to suggest different ways of selection. A list of concepts and sentences will help the person to choose a graph typology. I believe that moving the attention to the aim will help the person to solve his problem and understand the information. IBM's Many Eyes propose an approach in this direction classifying his graph under these areas: see relationships among data points, compare a set of values, track rises and falls over time, see the parts of a whole, analyze a text, see the world. This IBM experiment for data visualization

works online producing visualizations but also recording and publishing all the data submitted. In this experiment, the chosen verbs are comparison, composition, distribution, relation and trend. Each one will offer a selection of diagrams that could be used for the purpose. For example a Pie chart for a composition, or a classic column graph for a comparison. In this first attempt the offered diagrams and charts are basic, but due to the possibilities of expansion of the open source platform and the availability of examples from designer that use Processing as tool for their data visualization is not hard to believe that a wider implementation can be achieved.

Results. The third and last part of the screen shows the result. Before the selection of the chosen graph, the gallery of graphs and charts will filter his content according to the purpose selected. After the selection, the entire portion will show the graph allowing some customization. At this point, a small bar will appear on the last part of the screen on the right allowing saving and exports the file.

The result is a vector file that can be used as it is or integrated in a wider workflow by designers and professionals. A toolkit designed for common people don't need to have limitation in usage by advanced and skilled users.

3 Conclusions

In this study, I tried to reach an operative definition of Small Data. This emerging concept seems to achieve a good acceptance through media, and I believe can be part in the process of creating and enhancing a creative mindset in opposition to the consumptive mindset [14]. The project proposal, working on the same direction, tries to propose to designers a broader way of conceiving toolkits that can be generalized to other artifacts. The early stage of development and testing of the toolkit, cannot allow a more critic position, but I believe that as open source platform the development can be wide. A future development can involve the number of charts available but also a mobile version for tablets.

In this study the restricted area of experimentation allowed a more theoretical reflection on data and the relation with people; a further analysis will be more carefully point at the successful and unsuccessful choice of the toolkit design.

This study starting from a preliminary example wants to open some general perspectives for other designers and developers. I think that the role of communication designers is to spread information and support people. Designers should work with contemporary tools and regarding contemporary needs to enhance people emancipation.

Figures

Fig. 1. An area chart with its relative code in JavaScript InfoVis Toolkit. In the upper right the data loading, followed by the code for the visualization.

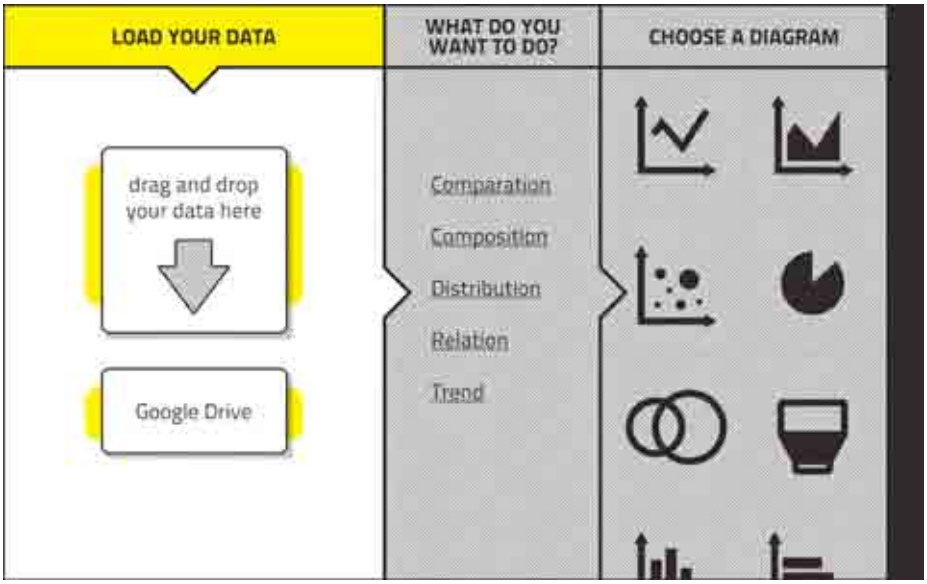


Fig. 2. The opening screen of the Small Data Visualization Toolkit.

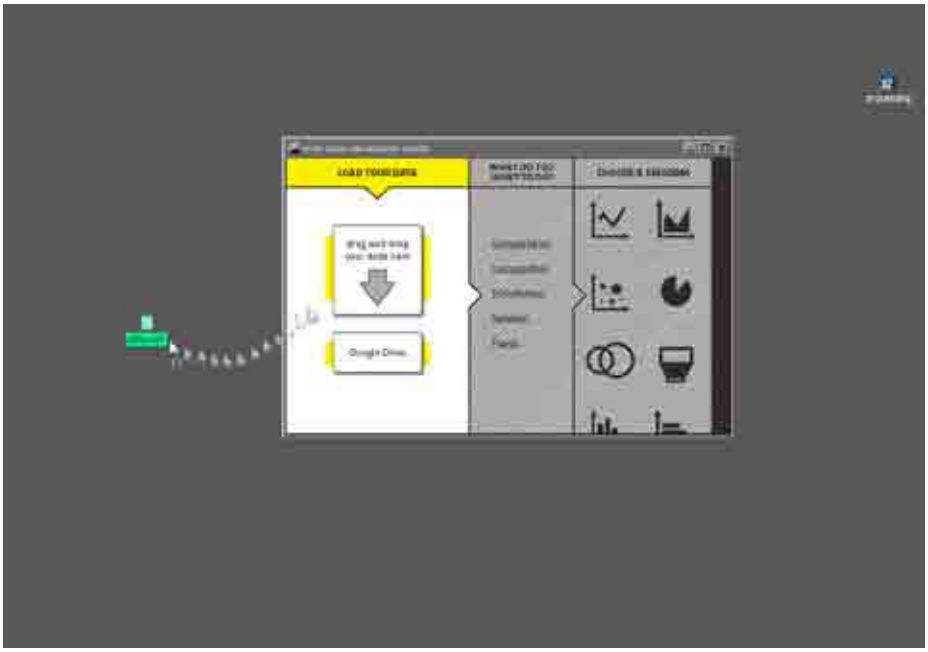


Fig. 3. Loading a data set from file in Small Data Visualization Toolkit.

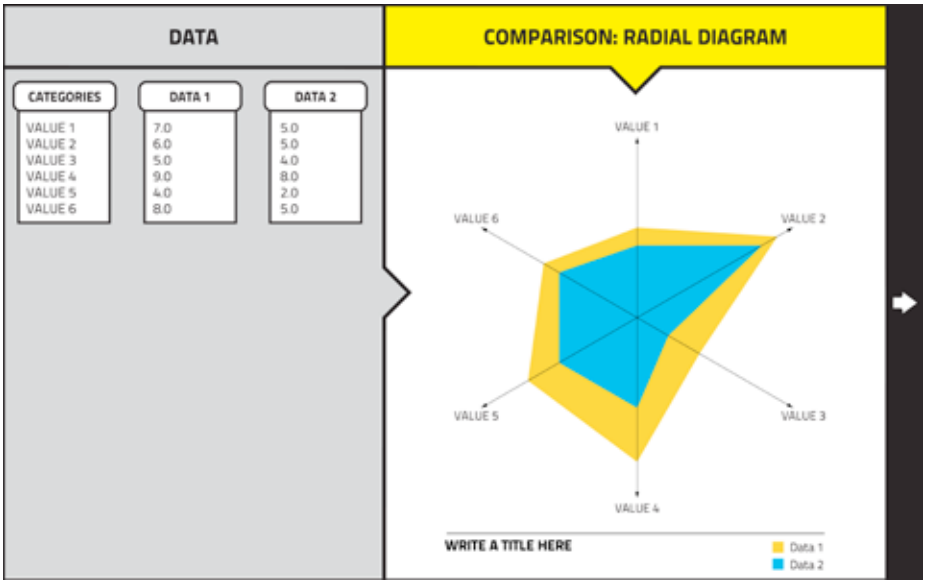


Fig. 4. A finished graph in Small Data Visualization Toolkit, ready to be saved.

References

1. Zicari, R.V.: Big Data for Good: A distinguished panel of experts discuss how Big Data can be used to create Social Capital. (June 5, 2012). Available at: http://www.odbms.org/blog/2012/06/big-data-for-good/ (September 25, 2013).
2. Mayer-Schönberger, V., Cukier, K.: Big Data: A revolution that will transform how we live, work and think. John Murray, London (2013).
3. Sankar S.: The rise of human-computer cooperation. (September 8, 2012) [Video file]. Retrieved from: http://www.ted.com/talks/shyam_sankar_the_rise_of_human_computer_cooperation.html
4. Turckle, S.: Alone Together: Why We Expect More from Technology and Less from Each Other. Basic Books, New York (2011).
5. Gershenfeld, N.: When Things Start to Think. Henry Holt and Co., New York (1999).
6. Lohr, S.: The Age of Big Data (February 11, 2012). Available at: http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html
7. Laney, D.: 3D Data Management: Controlling Data Volume, Velocity, and Variety. (Feb 6, 2001). Available at: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf
8. IBM, The Four V's of Big Data (access September 1, 2013). Available at: http://www.ibmbigdatahub.com/infographic/four-vs-big-data
9. Randall, C., Big data, small data (May 29, 2013). Available at: http://www.marketingweek.co.uk/strategies-and-tactics/big-data-small-data/4006583.article
10. Warren, J.: The Promise of "Small Data" (July 17, 2013). Available at: http://techpresident.com/news/24176/backchannel-promise-small-data
11. Pollock, R., What Do We Mean By Small Data (April 26, 2013) Available at: http://blog.okfn.org/2013/04/26/what-do-we-mean-by-small-data/
12. Estrin, D.: What happens when each patient becomes their own "universe" of unique medical data? (April 18, 2013). [Video file]. Retrieved from: http://www.youtube.com/watch?v=lAEhSGYEHWU
13. Bonsiepe, G.: Dall'oggetto all'interfaccia: mutazioni del design. Feltrinelli, Milano (1995).
14. Sanders, E.B.-N.: Scaffolds for Building Everyday Creativity. In: Frascara, J. (eds.) Designing Effective Communications, pp. 65—77. Allworth Press, New York (2006).

Relation Visualization for Semantically Enriched Web Content

Dimitris Spiliotopoulos[1], Ruben Bouwmeester[2], Dimitris Proios[3]

[1] Innovation Lab, Athens Technology Centre, Greece
d.spiliotopoulos@atc.gr
[2] New Media, Innovation, Deutsche Welle, Bonn, Germany
ruben.bouwmeester@dw.de
[3] Dept. Information and Telematics, Harokopio University of Athens, Greece
dimi.proios@gmail.com

Abstract. Many methods for web archiving include semantic analysis of the archived content. Texts from social media provide a rich feature set of analyzed data, such as topics, opinions, events, and more. One of the main challenges is the successful utilization of the semantic and raw data for an optimized search and retrieval approach. This work reports on the user needs for semantic relation visualization and the experimental approaches for creating visualizations that answer to specific user queries, based on the user interaction monitored by a dedicated semantic search tool.

Keywords: Semantic analysis / Twitter data / visualization

1 Introduction

Social media is a source of abundant user-created information that exhibits certain characteristics. The content is diverse, multimodal, opinionated, and can be classified according to popularity, influence, and other social factors. Current applications try to harvest social media content and present aggregated results to users. For Twitter data, there exist many applications that analyze and visualize sets of data according to user location information, Twitter post language, hashtag mentions, user mentions, post volume over time, keyword tag clouds, top X influencers, sources, URLs. The data that are used are meta-information already provided by the Twitter API and can be used for a helpful quick overview of statistics from data samples.

Recent approaches, mainly in web archiving, include social media content collection and analysis. Some of them, such as the ARCOMEM[1] approach, include a thorough post-level analysis on the semantic level of the content, identifying persons, locations, organizations, opinions expressed on them, topics discussed and events identified, clusters of semantically related entities, hashtags and users, and so on [1,2]. Applications may use the semantic meta-information to an-

[1] ARCHive COmmunities MEMories, www.arcomem.eu

swer user queries such as "return all positive opinions on President Obama" that can also, with the use of faceting, be refined to "return all positive opinions on Obama from European politicians on the economic crisis". Such queries are concrete and clearly formulated to retrieve web and social web data that match one or more parameters, either from raw data or semantically analyzed meta-data. Results can be visualized in the form of bubble graphs and timelines using frameworks such as D3 and present the current common practices [3].

Semantic analysis is a key feature in web archiving. In the cases where social media content is also archived and analyzed, the abundance and complexity of the semantic information provides a key advantage for retrieval tasks as well as high complexity for user interaction. Semantic-centric approaches to data visualization may provide a direct mapping between user and data semantics [4]. That can be used to guide navigation over large collections of documents through semantic visualization [5]. However, one of the most important issues is the unobtrusive integration with the user interaction [6].

Semantic analysis has also enabled the design of semantic search and retrieval methods and applications that utilize the semantic information to a great extent [7]. Such approaches provide results that combine raw data (such as hashtags and users), semantic data (such as opinions and topics) and statistical figures on combinations of the above. This work examines two case studies. The first uses the #BostonMarathon #Bombings of 2013 Twitter data and examines how specific data may stand out during visual exploration. The second uses the US Election 2012 web archives and the ways of building semantically related content visualizations that can be used to display search results that correspond to semantic queries such as "show me everything on Obama and Romney" and follow up.

This section provided the introduction, motivation and related work while the next sections discuss the social media and user interaction driven visualization problem, experimentation, implication and methodology for search and retrieval from big data archives. Finally, conclusions and further work are presented.

3 Twitter Data Experimentation

The first part of the work was aimed at establishing a base of user understanding on the importance of Twitter specific features, such as hashtags, with topics and sentiment. The generic problem that is faced in this domain is that Twitter posts are by default too small to extract topic information. As such, it is impossible to align Twitter data with topics identified in the web documents.

In the search and retrieval tasks, accuracy is of paramount importance, so specific solutions must be applied in order to associate Twitter data with topics, entities and sentiment values. One way would be to group similar Twitter posts and treat them together as a web document. The hashtag information can be used in this set up. However, this could result in large numbers of posts under each hashtag, manifesting itself in the visualization of the search results, since it would be impossible to return all data from one or more hashtags. To add to that,

the nature itself of the Twitter data, governs the expectations of the users. In that respect, the users expect to be able to retrieve keywords/catchwords, opinions over time frames of events, events that the posts refer to, influence information, in effect a reply to the question of what have people been talking about regarding an event.

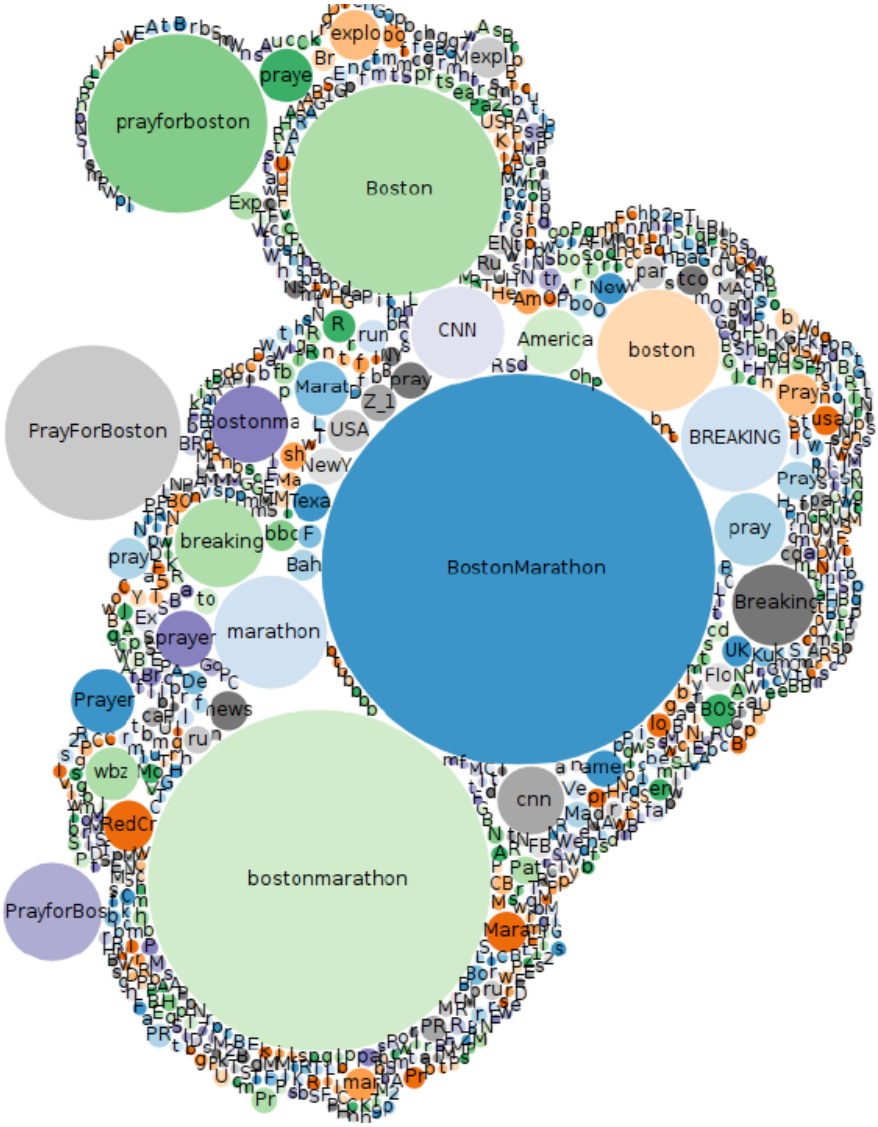


Fig. 1. General Hashtags Bubble Graph

The dataset used for this exercise contained 5000 tweets related to the Marathon event and the bombings that took place in Boston on 15/4/2013. The Mara-

thon started at 09:00 and the explosions occurred at 14:59. The tweets were collected in the timeframe between 14/4/2013 at 21:00 and 15/04/2013 at 19:46. Each of these tweets was annotated for polarity (positive, negative, neutral) according to its content. Some examples of the tweets are:

a) Excited for the #bostonmarathon tomorrow #makeitcount #findgreatness
b) Best of luck to all those running #bostonmarathon today! Have FUN and enjoy!!
c) #bostonmarathon explosions! Horrifying site :(let's pray for the may affected
d) #PrayersforBoston our prayers go out to Boston this afternoon.

With the above considerations, two types of data graph visualizations were considered: the hashtag graph and the sentiment chart. Since topic detection for single tweets is not recommended, hashtags could be used for the top-level visualization. The main tool that was used to create the various implemented graphs is JavaScript and more specifically the d3.js JavaScript library [3]. D3 is used for manipulating database documents using HTML, CSS and SVG components displaying JSON-formatted data.

As seen in figure 1, each topic may correspond to a set of twitter hashtags whose names may exhibit similarities as close as very minor lexical or stylistic transformations (e.g. “BostonMarathon”, “bostonmarathon”). These hashtags are detected using simple heuristics or Levenshtein distance [8] and are consolidated as shown in figure 2.

Fig. 2. Bubble Chart of consolidated hashtags using logarithmic scale for radial size

Sentiment charts are also one of the standard requirements. Clicking on the hashtag of interest can access them. They present the frequency of the positive and negative tweets over time. As shown in figure 3 below, the number of tweets for the Boston Marathon was relatively small in the beginning, however, after the bomb explosion it was rapidly increased. As was also shown, the negative tweets dominated over the positive ones as time was passing, since more people expressed their sadness or anger about the event. The fact that many positive tweets are detected (as shown in the graph) is mainly due to many people expressing hopes and wishes (e.g. "Best wishes to those at #BostonMarathon", "I hope everyone is ok").

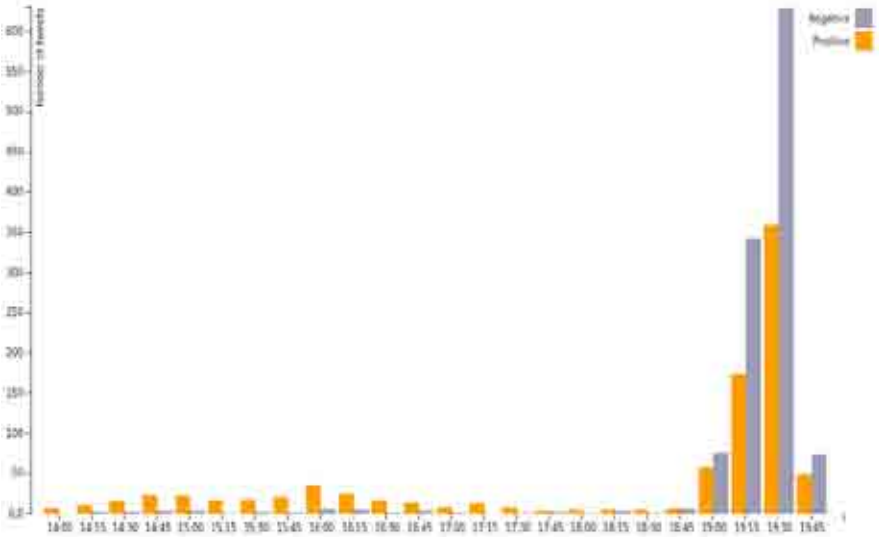


Fig. 3. The distribution of positive and negative tweets per hour

4 Semantic Relation Visualization

For the purpose of this task the Search and Retrieval Application (SARA) was used to measure user interaction [4]. As part of the heuristic evaluation, the users were asked to search using both a scenario that was designed to cover all functionalities and features of the SARA interface and as free search using any means necessary to complete their tasks. Their interaction was automatically logged and their intentions were recorded using the think aloud method.

The archived web documents were indexed in Solr [9] as semantic data including named entities, opinions, events, topics, as well as post-processed entries such as clustered entities. One of the major findings of the users was the fact that, after the initial search by one or more entities of interest, filtering by topic is the optimal way of retrieving the required information. However, only the

nearest topics are suggested since that process is search-specific. Follow up queries on topics and information contained in the topics has to be explored. This can be remedied by visualizing the map of the detected topics and, even better, as a follow-up requirement, provide a high level visual of all topics for a whole set of collected resources. Figure 4 shows the top level visual regarding topics. The distance of the topics between them is provided by the layout itself while the number of contained web resources by size of each topic bubble.

Fig. 4. Topic influence display

The users may click on the topics to explore them. By doing so, the major associated entities of the chosen topic can be viewed as well as indicators on detected opinions on each, as shown in figure 5. That, in effect, projects the overview of the entities of a topic that span all the web resources of that topic. At that point an entity may be clicked or the topic in order to zoom out to the figure 4 overview. By choosing an entity, all the topics that contain that entity as a major entry can be viewed as in figure 6. In effect, the users may always go forward in their search by either “zooming out” to topic view or “zooming in” to entity/opinion level.

Fig. 5. Topic, entities and opinions

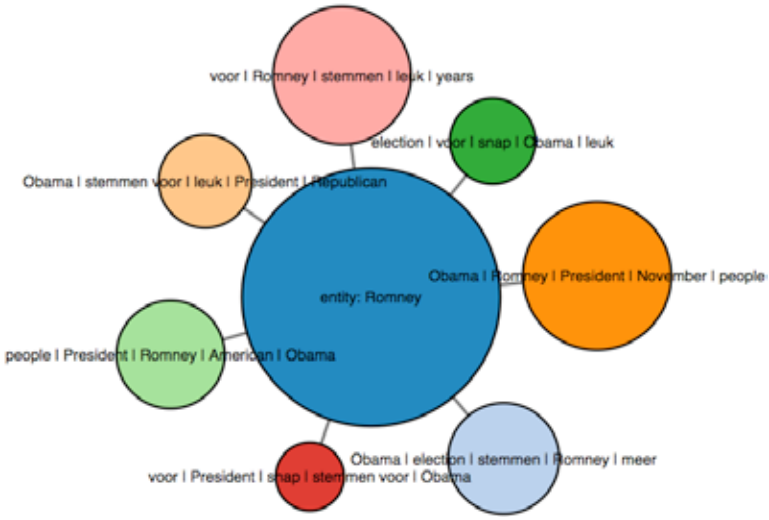


Fig. 6. Entity and relevant topics

5 Evaluation

The US Election 2012 dataset was used for the evaluation. The interactive visualizations were made available through the SARA web interface from the start of the search process. The purpose of the usability evaluation, at that point, was to measure user engagement with the newly introduced visualizations, changes in the key performance indicators regarding the search and retrieval core tasks, verification of the findings of the heuristic evaluation mentioned in section 4, and evaluation of impact and acceptance by the end users on the interaction level.

The typical scenarios for search and retrieval used in earlier iterations of usability evaluations were slightly adapted to include optional use of the topic-based visualizations. The hypotheses for the above were:

i. Taking into account the clear preference that users have shown to the introduction of topics in the ARCOMEM data and respective semantic search functionality, it was expected that the end users would actively show clear preference to using the interactive visualization process. The drawback is that using the visualization exclusively would yield results for certain typical search expectations, but it was not designed to single-handedly replace the traditional features of such application. For example, the user cannot filter or facet the results, by other entities, source network, nor rank the results according to opinions. Furthermore, many users may search by topic or by entity, entity being the norm for non-experts or non-archivists.

ii. There was already a significant improvement in accuracy and minimization of exploratory or corrective backtracking with the introduction of topics. Given the fact that, by design, the visualizations provide an accurate overall view on topics, it was expected that the indicative times to reach the same state would be even shorter.

iii. The visualizations were fine-tuned on the interaction level, based on the heuristic evaluation that was also used to establish the baselines for visual interaction. The findings were expected to verify the previous conclusions, however, this time around it was the end users of all levels that participated.

iv. The impact and acceptance was expected to be higher than the text-only interaction that was evaluated at earlier stages. Items and actions of interest were flagged for further investigation such as the topics/entities/opinions relation view (figure 5) regarding over-information and understandability.

The feedback from the users indicated that the topic-centered approach was welcomed and accepted by most users, however, an additional request was to provide a higher-level relation of entities and topics at the start of the search. This was expected, since it is common, especially for targeted search, to use entities for searching archives and move to more abstract level, such as topics, later if further search is required. A simple response to that was to include an entity word cloud (figure 7) at the same time, connecting the major entities of the web documents to other entities (i.e. clustering) or topics by colour.

The speed and accuracy were also improved since the visualizations allowed the users to engage in topics and described entities is just two clicks while, pre-

viously, topics were made available only during the search, whether at web resource level (lowest) or as prediction based on the interaction. The users now felt that they had more choice, especially true for the experts.

Fig. 7. Entity word cloud, clustering is indicated by matched colour

For the topic/entity/opinion view, some users had difficulty scanning the graph. There is clear feedback that the type of entity should be shown, preferably by color while only the most opinion-heavy entities should be visualized, resulting in less clutter. On the interaction level, the process was deemed highly acceptable. It was also mentioned by a large percentage of the participants that the social media results should be also visualized as part of a comparative view with the traditional web results or aggregated.

6 Conclusions and Further Work

The idea behind this work was not to build an ontology visualization toolset [10] but provide an interaction-aware interactive visual way for optimised semantic search and retrieval. Users perceive the potential of interactive visualized content by making semantic inferences [11] and the main task was to meet those efficiently on the interaction level. The described approach does not require a query formulation by the user at the start of the search and retrieval process but rather automatically generates the queries needed during the interaction.

Further work includes experimentation on a combined semantic relation visualization of social media meta-information, such as hashtags, with higher level semantic descriptions such as topics. That is expected to result in

realizations of summative results, like topic-ver-time, topic influence over hashtags, and so on. It is also expected that nested levels of topics or other interemediate semantic levels may be expressed, especially for very large collections and broader domains.

Acknowledgements

The work described here was partially supported by the EU ICT research project ARCOMEM: Archive Communities Memories, www.arcomem.eu, FP7-ICT-270239.

References

1. Risse, T., Peters, W., Senellart, P.: The ARCOMEM Approach for Social and Semantic Driven Web Archiving, In: Proc. 1st Int. Workshop on Archiving Community Memories, Lisbon, Portugal (2013)
2. Schefbeck, G., Spiliotopoulos, D., Risse, T.: The Recent Challenge in Web Archiving: Archiving the Social Web. In: Int. Council on Archives Congress, Brisbane, Australia (2012)
3 Data-Driven Documents, http://d3js.org/
4. Chen, C., & Carr, L.: A semantic-centric approach to information visualization. In: IEEE Int. Conf. on Information Visualization (1999)
5. Kboubi, F., Chaibi, A. H., BenAhmed, M.: Semantic visualization and navigation in textual corpus. Int. J. Inf. Sciences and Techniques (IJIST) Vol.2, No.1 (2012)
6. Voigt, M., Pietschmann, S., Meißner, K. Towards a semantics-based, end-user-centered information visualization process. In: 3rd Int. Workshop on Semantic Models for Adaptive Interactive Systems (2012)
7. Spiliotopoulos, D, Tzoannos, E, Cabulea, C., Frey, D.: Digital Archives: Semantic Search and Retrieval, In: Holzinger, A., Pasi, G. (eds.) HCI-KDD 2013, LNCS 7947, pp. 173-182 (2013)
8. Levenshtein, V.I. Binary codes capable of correcting deletions, insertions, and reversals. Technical Report 8 (1966)
9 Apache Solr, http://lucene.apache.org/solr/
10. Temitayo, F., Stephen, O., Abimbola, A.: RROVT: A Proposed Visualization Tool for Semantic Web Technologies. J. Information Engineering and Applications, 2(3), 7-25 (2012)
11. Keller, R.M., Hall, D.R.: Developing Visualization Techniques for Semantics-based Information Networks. In: ACM Workshop on Visualization in Knowledge Engineering, 2nd Int. Conf. on Knowledge Capture (2003)

Analysis and Visualization of Personal Data: the Value of Images in Social Networks

Pietro Costa[1] and Michele Zannoni[2]

[1] IUAV University of Venice, Italy
pietro.costa81@gmail.com

[2] Università degli Studi della Repubblica di San Marino,
Department of Economics and Technology, Republic of San Marino
michele.zannoni@unirsm.sm

Abstract. The goal of this research project is the analysis and development of tools to highlight and interpret personal data that is freely shared on the social networks of the web. Starting from the premise that images taken by mobile phones can be considered to all effects personal data representing values that can be displayed through a system of interactive maps, we will develop a strategy to identify the means for comparing different sets of data in order to reveal temporal and geographical information. The current project is an online platform for collecting and displaying interpolated data obtained by cross-reading the social networks Twitter and Instagram. The comparison between tweets and the pictures shared on Instagram will find some correlations that will be read and interpreted through an interactive interface. The results will be visualised on the basis of their hashtags and geolocation. Given a defined time interval, a geographical reference located in metropolitan areas, and a matrix of different hashtags, the software will create interactive maps by collecting different sets of data. This will provide the user with a tool for analysing and interpreting certain complex phenomena related to certain specific sociocultural aspects. The selection of case studies for the research has been carried out mainly within the social context that revolves around fitness. The underlying reason is that the use of camera images is strongly linked to personal motivation and self-confirmation. The correlation between the satisfaction of achieving a specific performance in sports (though not a professional one) and the temporal and geo-referenced context becomes a tool to analyse and understand complex dynamics evolving within a multifaceted environment such as the one that orbits around sharing online pictures of the results as self confirmation.

Keywords: personal data / interactive maps / hashtag

1 Introduction

In recent years there has been a growing trend in the collection and use of the increasingly ubiquitous *Big Data*; as a consequence it is becoming more and more important to analyse this flow of data, drawn from a growing variety of sources. This data is universally collected and recorded: sources of "production" might include, for example, sensors for collecting information of various kinds, GPS and RFID signals, posts on social media websites, digital videos and images, commercial and financial transactions.

In particular, major Internet companies such as Google, Amazon or Facebook, are interested in collecting and extracting vast quantities of personal data about people and their activities. Individual users are also active in the digital recording and storage of events and personal experiences thanks to personal informatics tools [1], systems for monitoring and collecting personal data that help the individual user track his everyday life from various psychophysical points of view. Recently, a growing number of applications and online services have been developed that can offer direct support to users thanks to the data they diffuse publicly: examples might include the real-time roadway traffic conditions provided by Waze[1] or the social network based on geolocalization, Foursquare[2].

As stated by Viktor Mayer-Schonberger and Kenneth Cukier [2], we have now entered the era of "datafication", understood as the quantification and relative transformation of reality into data: we have shifted from understanding the causality of events to correlating and predicting them. Because these data flows are acquired and generated by many different sources, they appear structured in different ways. To be able to use them, it is essential to design effective ways of selecting, elaborating and presenting the mass of available data on time, in terms of both quantity and quality: an exercise in interpretation that plays a primary role in the processes of interaction with specific contexts.

2 Areas and methodologies of the research study

This research study may be framed within the context of interface design; its ultimate goal is to design a tool to collect, compare and visualize data and information that may be of use to the user. Specifically, it will be interesting to explore how the design of this web-based tool may allow non-specialized users to focus search queries finalized towards acquiring data and making it easier to understand. For example, while individual users of Instagram rely on the application merely to store and share images, their collective activity reveals an astounding quantity of geographical and visual information on the world. The above-mentioned tool would make it possible to search this type of information, chang-

1 http://www.waze.com/
2 https://foursquare.com/

ing the approach to searching complex data and improving the tools available to users for their visualization.

In the process of searching for the sources of the data, we tried to identify the most appropriate social networks for acquiring shared data, not influenced by the way the connections between users are managed. During the initial selection phase, in addition to Instagram and Twitter, we had also included Facebook, but from the analysis of its API we realized that it was forbidden to read user data, even anonymously, for obvious reasons of privacy. For the same reason, in 2010 Facebook forced an earlier very similar project created by Pete Warden[3] to shut down, because it broke the rules for using the social network, and above all it was unclear how the data was acquired. Whereas in the case of Instagram and Twitter, because of the very nature of the two social networks, the data is public: only Instagram later allows the user to decide to protect his images.

Social media are being studied by many disciplines and from many different points of view, with specific attention to the analysis and visualization of large groups of digital images and the relative geo-spatial and temporal metadata they contain. On this subject, see the recent project entitled Phototrails[4] [3, 4] on the visualization of millions of images captured by Instagram, or the considerations on tracing behavioural patterns by associating geo-referenced visual data [5, 6].

3 The technological context: the rules and limits of social network APIs

The concept of open platform and its capillary diffusion in recent years has led to greater communication between social platforms and a multitude of web services that have gradually been built into many areas of communication.

The importance of the value of social data, consistently underlined in many analyses conducted by economists [1], emphasizes how personal data is increasingly becoming an economic asset for the major social networks. From a technical point of view, the integration has been made easier by the constant development of new APIs (Application Programming Interface) that simplify software integration by standardizing the data between the various social platforms.

To search and store the data, the tools offered by the social networks are not all alike in their services and calls to the server. The heterogeneity of these methods makes it impossible to complete all types of searches on all the social networks. In the case of Instagram and Twitter, all the data is public and may be searched by default. In other social networks such as Facebook, personal data is not public, in order to guarantee privacy. The data owned by Facebook is available to individual companies but may not be consulted freely.

---

3 Pete Warden, "How to Split Up the US" http://petewarden.com/2010/02/06/how-to-split-up-the-us/

4 http://phototrails.net/

4 The project and project phases

During the development process of the project, several working tools were conceived and created to guarantee a different approach to the understanding of personal data freely available on the web.

The first tool to be created was a system for acquiring data stored on an online server. This software, which was developed using the APIs of Instagram and Twitter and other open-source libraries, serves to collect and record the data on a dedicated database. In the case of Instagram, the server downloads and temporarily records a smaller format of the image, so that it can be analyzed in terms of possible dominant colours and in its histograms. Once the data is saved, homogeneous values are selected and recorded in the fields of a spreadsheet of the database. For each element downloaded from Internet, the original text format is also saved for possible data elaboration in the future. The decision to preserve the raw data was made to guarantee that in the future, a specific phenomenon can be analyzed with new processes and different interpretations.

There are several existing projects similar to ours, ranging from commercial to consumer applications. One of the most famous and freely accessible services is *Statigr.am*[5], a research engine based on the Instagram API, which allows searching by hashtag. In the case of Twitter, many websites support various search modes, like *Twitterfall*[6], for example, which tracks tweets in real-time by hashtag and location. Other tools such as *Hashtags.org*[7] and *Tagboard*[8] are research engines that perform wide-ranging searches by hashtag on multiple social networks. There are also open-source projects such as *QIS (Quick Instagram Search)*[9] by Al Shaw, which have developed advanced research routines in Instagram; by using this algorithm, the author was able to find pictures taken at the site of the Boston Marathon bombing in 2013, just a few minutes before the explosions went off. *Datasift*[10] is a commercial alternative founded in 2010 that stores the activities of multiple social networks on its servers, and allows them to be searched for a fee.

The analysis conducted during phase one of the project showed that many of these services are generally limited in their approach to searching: they only search small quantities of data and make no attempt at visualizing them rationally to offer users instruments of analysis.

The project, renamed *ReverseHashtag Project* for this phase of experimentation, is developed on three levels: a tool for acquisition (Fig. 1), one for data post-processing, and one for visualization. Each of these tools represents an element of the design process in the development of the project.

5 http://statigr.am/
6 http://twitterfall.com/
7 http://www.hashtags.org/
8 http://tagboard.com/
9 https://github.com/propublica/qis/
10 http://datasift.com/

4.1 How data is searched on social networks

The *ReverseHashtag Project* is based on minimal units of acquisition defined as "nodes". Each node is an action that defines a query to a data source. For example an acquisition node could be a query to Instagram based on a specific hashtag. Another possible research typology on Instagram could relate to specific geographical coordinates, to a specific date, or to a combination of these elements. Each phase of data collection is a delicate process of calibration and as such requires the visual feedback provided by the interface. Before being stored, the acquisition is interpreted as unprocessed raw data.

Fig. 1. User interface of the tool for acquisition during the storage phase.

During the storage phase, users become anonymous to protect their privacy and to respect the legal terms of the licenses to use the two social networks.

During the testing phase of the process, a data collection project was initiated to record the activity of the social networks Instagram and Twitter inside the Luzhniki Stadium in Moscow during the World Athletics Championships in 2013. Approximately 11,000 elements were acquired from Instagram and 1800 messages from Twitter. On the closing night of the event alone, 1875 photos were posted and 237 tweets were sent. To explain this difference, it is important to understand the nature and the intrinsic characteristics of the data, which in the case of a photograph is easily motivated by the user's desire to document himself and the sports event.

The interface for managing the data is designed to be used by either a common browser or a mobile application for real-time data collection.

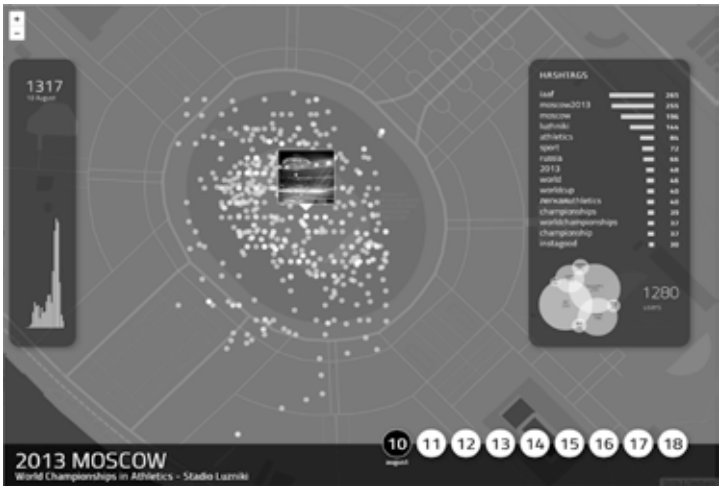


Fig. 2. The activity of the social networks Instagram inside the Luzhniki Stadium in Moscow during the World Athletics Championships in 2013.

4.2 The construction of derived data: how the elements are rendered homogeneous

Every collection process requires a further phase of elaboration to acquire derived data, which can be obtained by interfacing the system with other APIs, to localize the data with greater precision for example, or to assess the evolution of given users.

When a tweet or an image is shared online and is freely accessible, it becomes part of our reality, which for the most part takes place within a precise geographical location. This simple action, which may appear irrelevant in the long term, constitutes an invisible but real complexity. To understand and study this complexity requires the creation of tools that can bring it to light and build logical routines to transform the data, making it homogeneous and suitable for comparative visualizations. Extremely interesting contributions may be found on the theme of elaborating data from photographic images in Lev Manovich's research studies [7]: with his experimental software *Imageplot* he searched correlated data in the information contained in the images themselves in terms of chromatic dominants and colour thresholds.

Reading complex data in business has always been a specific task performed by professional figures, trained to interpret values that in most cases carry no significance for ordinary people. These figures, who are capable of understanding the complexity of the data, are becoming increasingly important in many strategic and decision-making areas. This makes it necessary to create tools to simplify the association and comparison of complex data for people who are not experts in this field.

To achieve this result, the data in the *ReverseHashtag Project* must be reconstructed and integrated before being visualized. This process of enrichment and of simplification at the same time, is based on a continuous process of re-elaboration, which takes place both during the acquisition of the data and in later post-editing processes.

4.3 How the data is visualized

The final phase of the process in *ReverseHashtag Project* is the visualization of data collection. This is a critical issue in reading the data, which can go in two possible directions: the first is to comply with the visualization presets defined by the project, the second is to export the data in a format that can be used by the most common visualization platforms in exchange formats such as CSV (comma-separated values) or JSON (JavaScript Object Notation) to create a visualization by means of an interface developed in HTML5.

In either case, every dataset obtained in a specific collection process, for obvious legal reasons, cannot become public because this would violate Instagram's user agreement, but the conceptual maps and the resulting visualizations could on the contrary be freely published as the result of an effort to interpret public information freely accessible on the web.

To guarantee adequate support for the visualization of the data, an analysis was conducted on case studies for the representation of photography-related personal data, with the purpose of defining which presets for the use of the data would be most appropriate for this project.

The results show that three principal tools are required to understand the collection process in the initial phase: the first is a geo-referenced representation (Fig. 3), the second will be based on the curve of hashtag repetitions, and the third on linearly-distributed chronological elements. The integration and co-existence of these three presets should make it possible at any moment to create increasingly complex visualizations, to progressively simplify understanding.

Fig. 3. Instarunners community: geo-referenced representation of users activities.

5 Case studies

Following an initial testing phase on various locations within the project area, the decision was made to apply the *ReverseHashtag Project* to a first case study in the field of outdoor fitness.

There is a large Instagram community linked to the hashtag #instarunners with over 120,000 shared photographs of which 40,000 were uploaded over the past year (Fig. 4). Within the group of data tied to a specific hashtag there is, depending on the cases, a certain quantity of elements that that are actually a type of advertising, often devoid of GPS coordinates and associated to too many tags. By filtering the results it is possible to obtain a sample of data that plausibly represents real people. First, the totality of the data was represented on a map in an attempt to understand the geographical distribution of the individual photographs. At a later phase the visualization tools were fine-tuned to identify words or hashtags that recur in the contents of the items.

Fig. 4. Instarunners community: chronological distribution of the elements.

The time factor in relation to this type of community is very important: filtering and representing the data based on the element of time, can help to understand the habits and changes in the behaviour of the users over an established length of time. The repetition of a ritual action, such as posting an image, becomes a motivating element in which the users seek confirmation from their own community [8]. This complex phenomenon currently has no forms of representation because, in the mass of information, the semantic content cannot be interpreted and the only action that an analysis software algorithm is capable of is cross-referencing and highlighting the repetition of specific words or hashtags. The final visualization that will emerge from the case study will highlight, in relation to a global context, the habits of a specific community of people who practice

outdoor fitness (Fig. 5). Our approach to the understanding and visualization of this specific reality was to design a tool that could obtain comparable data which could be represented and used to study and observe this complex phenomenon.

Another application of the project associated with the world of fitness was performed by running a search on the tag #running on Twitter and analysing the results over a week's time. This case study is still in the elaboration phase and its goal is to investigate the user's daily routine, in order to verify whether the project's visualization models can provide an exhaustive representation of the changes in the users' training habits over the course of the week.

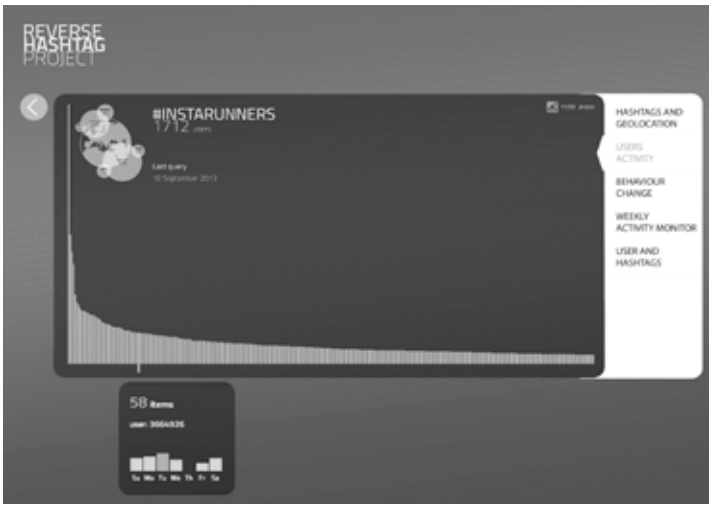


Fig. 5. Instarunners community: visualization of users activities.

6 Future developments of the project

The *ReverseHashtag Project* is currently in an experimental phase that requires a series of case studies to complete the development of the model of representation and the visual tools for the analysis of the evolution of the individual users. Another field in which to pursue experimentation is the development of a model for acquiring data in relation to complex locations not circumscribed within a localized area, but along an itinerary such as a marathon route, for example.

A further development of the project will involve the model used to collect the data, with a change in the logic for recording the elements. The current model reads the history stored in Twitter and Instagram, and made public through the API. To avoid any filters for reading the data that might have been created to support the technical management of the social network databases, a system will have to be developed to collect the tweets and photographs on Instagram in real-time streaming. This change in the acquisition mode will make it necessary to programme the data collection beforehand, but the results will be more complete than what the social networks' own research engines make it possible to acquire.

References

1. Li, I., Dey A., Forlizzi, J.: A stage-based model of personal informatics systems. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 557--566. ACM (2010)
2. Mayer-Schönberger, V., Cukier, K.: Big Data: A Revolution that Will Transform how We Live, Work, and Think.. Eamon Dolan/Houghton Mifflin Harcourt (2013)
3. Hochman, N., Manovich, L.: Zooming into an Instagram City: Reading the local through social media. In: First Monday, 18(7) (2013)
4. Hochman, N., Schwartz, R.: Visualizing instagram: Tracing cultural visual rhythms. In: Proc. The Workshop on Social Media Visualization (SocMedVis) in conjunction with The Sixth International AAAI Conference on Weblogs and Social Media (ICWSM-12), pp. 6--9. (2012)
5. Crandall, D. J., Backstrom, L., Huttenlocher, D., Kleinberg, J.: Mapping the world's photos. In: Proceedings of the 18th international conference on World wide web, pp. 761--770. ACM (2009)
6. Kisilevich, S., Krstajic, M., Keim, D., Andrienko, N., & Andrienko, G.: Event-Based Analysis of People's Activities and Behavior Using Flickr and Panoramio Geotagged Photo Collections. In: Information Visualisation (IV), 2010 14th International Conference, pp. 289--296. IEEE (2010)
7. Manovich, L.: Media visualization: Visual techniques for exploring large media collections. In: Media Studies Futures. ed. Kelly Gates. Blackwell (2012).
8. Kamal, N., Fels, S., & Ho, K.: Online social networks for personal informatics to promote positive health behavior. In: Proceedings of second ACM SIGMM workshop on Social media, pp. 47--52). ACM (2010)