INSIGHTFUL STORIES: THE RISE OF DATA JOURNALISM AND VISUALIZATION IN AMERICAN NEWSROOMS

By Alberto Cairo

Introduction and Justification of Interest

Data is the new buzzword in communication, business, marketing, and public relations. As I've pointed out in the past, citing research by International Data Corp (IDC), The Economist reported two years ago that the information generated in 2010 alone reached 1,200 exabytes, an amount equivalent to thousands of billions of issues of the venerable British magazine. The story added that the total amount of extant digital information totals several zettabytes.

In August 2010, Erich Schmidt, former CEO of Google, announced in a conference that between the beginning of time and 2003, humanity generated roughly five exabytes of data; whereas we now produce the same volume of bits every two days. “The information explosion is so profoundly larger than anyone ever thought,” said Schmidt. Five exabytes is more than two hundred thousand years of DVD-quality video. Not all that ‘information’ is what you would call ‘information’ in a colloquial conversation. Most of it is the product of automated processes and communications between computers, cell phones and other devices: nothing that a human brain can understand. But still. Information overabundance is a reality today.

In the 1970s, years before access to the Internet was universal Richard Saul Wurman, then a professor of architecture in North Carolina, predicted that the foreseeable information explosion would require the intervention of a new breed of professionals trained in organizing data and making sense of it. Wurman would later call these professionals information architects (WURMAN, 1997), although most of what he wrote could be a guiding light for contemporary journalists, the professionals who are supposedly in charge of helping citizens understand the world.

Starting in the second half of the twentieth century, certain news organizations, mainly in the US, stepped forward to meet the challenge of an environment in which huge databases are publicly released on a regular basis, a landscape in which information is not scarce anymore, quite the contrary. A journalist today should not be just a professional who looks for data that has public relevance, but who handles those data to extract meaning from them.

News organizations nowadays recognize the need of incorporating uncommon professional profiles to their staffs, such as statisticians, programmers, and database developers, who can deal with the flood of information caused by new laws of transparency put into effect in many countries. News organizations are not in the business of finding information anymore, but in the business of making sense of that information. They have finally embraced the old 1947 saying: "It is no longer enough to report the fact truthfully. It is now necessary to report the truth about the fact."

Latin America and Spain have been notoriously slow in catching up with these new trends in journalism. Today, no Spanish newspaper has a team of statistical-savy journalists who can manage data-based investigative stories. Besides, the graphic designers working in newsrooms are not trained in modern information visualization techniques (visual display of data and phenomena). According to some of the most renowned investigative journalists in Spain, the lack of knowledge about formal research methods among journalists is a considerable challenge that should be faced with more training and new hires. The key element in knowing where to begin, they add, is to identify cases of success in other countries and study them in depth. That is the core of this research proposal.

My goal is to make a comparative analysis of three media organizations from Costa Rica, Argentina, and Brazil, that embody the three stages in the development of a data journalism team; then, I propose to relate them to the...
North American tradition, represented in this proposal by The New York Times. What do they have in common? What distinguishes them in their approaches to the handling of data? What are the key components of those teams? What are their strengths and their weaknesses?

La Nación, from Costa Rica, represents the consolidated effort: a solid, 5-peole desk who has been successful throughout the years. The investigative desk at La Nación uncovered several corruption cases that led to the arrest of two former presidents of the country. However, the department has gone through changes recently. According to its head, Giannina Segnini: “It has been only during the last year that the unit started to focus on database journalism and data-driven journalism as our main output. Now, more than 70% of our stories are being generated by these techniques.” La Nación (same name) from Argentina, is the desk-in-construction, as it was born just two years ago, but has experienced great success. Estado de São Paulo, from Brazil, shows how a data journalism team looks like when it starts to form.

Research about data journalism teams is highly relevant. Despite the fact that data journalism is hardly a new phenomenon, it has been neglected in the academic literature as a secondary element in the current upheaval that news organizations are going through. As we are about to see, communication and journalism studies even lack a solid, clear theoretical framework to discuss data-based journalism, as some papers (ESTEBAN, 2012) suggest. There is a great need to define what data journalism really is, and to study how it has been practiced today.

State of the Question And Literature review: from computer-assisted reporting to data journalism

Originally (COX, 2000), computer-assisted reporting refered to any information gathering done with the help of computers. Computers started to be broadly adopted by newspaper newsrooms when USA Today was born, in 1982 (PRICHARD, 2007), although they had been in use by certain specialists way before that. According to Cox "the first actual instance of computer-assisting reporting was with the 1952 presidential election when CBS employed the Remington Rand UNIVAC to predict the outcome of the race between Eisenhower and Stevenson".

By the end of the 90s, the definition of computer-assisted reporting started to change. All journalists were using computers on a regular basis already, so the original explanation of what computer-assisted reporting is was too generic. Computer-assisted reporting was reframed first as “precision journalism” by Philip Meyer (1973, 2002). Meyer defined precision journalism as the use of social science quantitative and qualitative research methods and tools to gather and filter data for news stories.

Today, it is common in academic and professional forums to talk about “data journalism”, a third phase of computer-assisted reporting and precision journalism marked by the increasing availability of open databases provided by public and private institutions, and by the broad variety of low-cost software tools and programming languages that journalists can apply to create visualizations.

From some recent publications (V.V.A.A., 2012) it is possible to infer that a “data journalist” is something more than a “precision journalist”: precision journalists used to use databases and statistical techniques to make sense of data. Then, they summarized those data and presented the results to readers in traditional stories and tightly edited graphs. In a modern news publication, though, a data journalism team goes one step beyond that: it also creates meaningful databases, but then it doesn’t just use them to understand the story; it extracts the main points in the data, as old precision journalists did, but it also puts the data in the hands of the public, it lets the public navigate those data using modern information graphics and visualization techniques, such as complex charts, maps, and explanation diagrams (CAIRO, 2008, 2012). The data journalist is not anymore just a mediator between the data and the public, but a designer of the interfaces the public can use to make sense of the data.

As an example, this information graphic by The New York Times, showing a huge amount of data from the US Census Bureau (http://projects.nytimes.com/census/2010/explorer):
US newsrooms have had database editors for a long time, but those professionals have always worked in the backstage, as providers of insights and suggestions to other colleagues. In modern days, the landmark in bringing data journalists under the spotlight was the WikiLeaks war logs on Afghanistan-Iraq, and the US embassy cables (see STAR and KELLER, 2011), between July and November 2010. Several top news organizations, such as The New York Times, The Guardian, and Der Spiegel received “an enormous amount of classified US government communications”. As Bill Keller explained (2011), the coverage of the WikiLeaks papers would not have been possible without the participation of the team of database journalists led by Aron Pilhofer, and of the graphics department at The New York Times. They were the ones who created the tools to sort, filter, and search the huge amount of pages leaked by Julian Assange. Today, The New York Times, has more than 40 journalists who practice (full-time or part-time) data journalism and visualization (CAIRO, 2008 and 2012). This situation makes this newspaper the role model other organizations try to emulate. ROGERS (2011) bluntly explained why data journalism suddenly became a buzzword in newsrooms worldwide after the WikiLeaks revelations: “Is data journalism? Is it journalism to publish a raw database. Here, at last, is the definitive two-part answer: 1. Who cares? 2. I do hope my competitors waste their time arguing about this for as long as possible. Once you’ve had (...) WikiLeaks, the startling thing is that no-one asks those questions anymore. Instead, they want to know, “how we do it”?” My purpose in this study is to explain how smaller and worse-staffed media organizations in Latin America have started to “do it” — quite successfully— and what are are the main shortcomings and challenges they face in their path to mimicking what they consider the ideal data-journalism organization, The New York Times, at a different, maybe less ambitious scale.
Research questions and goals

My main question in this research project is: Considering that The New York Times is universally regarded as the main case of success in modern data journalism, are media organizations interested in developing data and visualization teams in the right path to emulate its structure, ideals, and dynamics?

Other possible, secondary questions that will be answered are:

• How is a data journalism team put together? What are the stages in its development are?

• How does a data journalism desk develop through time, what kind of professionals are its core, and what other profiles are incorporated down the road?

• How do data journalism and visualization teams adapt to particular conditions in each country?

• What are the main limitations smaller media organizations face in terms of producing data-based investigative journalism and visualization displays?

Methodology

To answer the proposed questions, the first step will be to clearly define the terminology to be used in the study and to devise a history of computer-assisted reporting, precision journalism, and data journalism (COX, 2000). This will give us the proper context to later analyze the work of the chosen subjects.

The second step will be an in-depth analysis of The New York Times newsroom in relation to its data-journalism projects. This will involve the following parts:

• A quantitative and qualitative analysis of the staff related to data journalism. This survey will include the following variables: number of professionals, roles they have in the newsroom, professional skills, training, and years in the organization. Combined and aggregated, these variables will provide a clear picture of the components of a modern data journalism and visualization desk that, at a different scale, could be translated to other media organizations.

• In-house observation of the newsroom, to be conducted in December 2012: I will spend three days observing how data-based investigative stories are developed in The New York Times.

• Semi structured interviews with at least five leaders in the newsroom, based on the on-site observations: Aron Pilhofer (head of computer-assisted reporting), Steve Duenes (head of information graphics and visualization), Matt Eriksson (deputy graphics director), and Amanda Cox (head of statistical-based visualizations), and David Barstow (investigative senior reporter). The goal of these interviews will be to complement the results of the previous survey, giving it some necessary qualitative context. The interviews will be conducted in the first three months of 2013.

Third step: To make the comparative analysis, my sample will consist of the three main Latin American news organizations who practice data journalism and visualization today: La Nación (Costa Rica), La Nación (Argentina), and Estado de Sào Paulo (Brazil). They represent three stages in the development of a data desk: the established team, the desk-in-progress, and the starting team. In the three cases, I will follow the same procedures outlined before for The New York Times:

• A quantitative and qualitative analysis of the staff related to data-journalism in each of those organizations, to be conducted in the second semester of 2012.
• I will make in-house stays to *Estado de São Paulo* in July 2012, with *La Nación* (Costa Rica) in August 2012, and with *La Nación* (Argentina) in the first semester of 2013.

• Semi structured interviews with the data journalism teams at the three organizations, based on the observations made in the previous step.

Fourth step: the main goal of the three previous step is to distill the main features of data journalism and visualization teams in the four organizations analyzed (the role model and the smaller desks), and to be able to compare them. In relation to the main question of the study: given that Latin American organizations consistently cite The New York Times as the model to be emulated when it comes to data journalism and visualization, what have they already done to follow that model and what are the features, professionals, skills, or newsroom practices and structures they still lack to achieve their goals.

**Notes and Bibliography**

1. Quoted from CAIRO (2012).


5. In contrast, the best North American news organizations do have designers with formal training in statistics, programming, etc. See CAIRO (2008).

6. Conversation (2011) with Antonio Rubio, head of the investigative team at *El Mundo*. Rubio is one of the promoters of a new Masters program in data-based investigative and data journalism: http://www.escuelaunidadeditorial.es/master-periodismo-de-investigacion/carta-director.html


10. Even keen observers of the transformations of new media, such as SALAVERRÍA (2008) devote little space in their studies to this booming area.


ESTEBAN, Chiqui. “Las nuevas profesiones del periodismo”. In El futuro del periodismo, Cuadernos de Comunicación EVOCA, 2012.


V.V.A.A. The Data Journalism Handbook. O'Reilly, 2012: http://datajournalismhandbook.org/