title: DATA-DRIVEN JOURNALISM AND THE PUBLIC GOOD

subtitle: “Computer-assisted-reporters” and “programmer-journalists” in Chicago

Sylvain Parasie (lead author)

University of Paris-Est Marne-la-Vallée, 5 bd Descartes, 77454 Marne la Vallée Cedex 2, France.

Email: sylvain.parasie@univ-paris-est.fr

Eric Dagiral

Paris Descartes University, 45, rue des Saints-Pères, 75270 Paris Cedex 06, France.

Email: eric.dagiral@parisdescartes.fr

Sylvain Parasie is an assistant professor of sociology at the University of Paris Est Marne-la-Vallée and a researcher at LATTS. His interests are in the area of media and technology, focusing on the sociopolitical implications of innovation in news organizations.

Eric Dagiral is an assistant professor of sociology at Paris Descartes University and a researcher at CERLIS. His interests are in the area of media and technology, dealing

with the role of producers and users in online innovation, especially in connection to quantification practices.

Abstract

Since the mid-2000s, some US and British news organizations have hired programmers to design data-driven news projects within the newsroom. But how does the rise of these "programmer-journalists", armed with their skills and technical artifacts, really affect the way journalism can contribute to the public good? Based on an empirical study in Chicago, we show in this article that although they have built on previous historical developments, these programmer-journalists have also partly challenged the epistemology conveyed by the computer-assisted reporting tradition in the US, grounded in the assumption that data can help journalists to set the political agenda through the disclosure of public issues. Involved in open source communities and open government advocacy, these programmers and their technical artifacts have conveyed challenging epistemological propositions that have been highly controversial in the journalism community.

Keywords

data-driven journalism, online news, investigative journalism, programmer-journalists, hackers, epistemology of news, computer-assisted reporting, database.

The role of computer scientists in journalism has increased significantly since the mid-2000s. Mostly in the United States and Great Britain, some news organizations have hired programmers - often calling themselves "programmer-journalists" - to produce innovative online news products (Royal, 2010; Daniel and Flew, 2010). Several leading newspapers (e.g. *The New York Times*, *The Guardian*), as well as independent news organizations (e.g. ProPublica) have set up dedicated teams within their newsrooms, specifically to design so-called "news applications". These innovative contents, consisting of online presentations, interactive maps or visualizations, rely on a wide range of computer techniques used to collect, process, analyze and visualize data sets.

Computerized data have been used since the late 1960s in American newspapers to support news production. In particular, the "computer-assisted reporting" tradition in the United States has encouraged journalists to investigate using computerized data. But several figures and scholars have claimed that contemporary "data-driven journalism" improves the way journalism can contribute to democracy – especially at a time when a growing number of data sets are released by governments. It is seen to do so mainly in three ways. The first is by strengthening journalistic objectivity. Tim Berners-Lee, for instance, claims that journalists "no longer find tips chatting with people in smoky bars", but have to "equip (themselves) with the tools to analyse (data)" in order to "help people out by really seeing where it all fits together, and what's going on in the country" (Arthur, 2010). The second way is by offering new tools to news organizations

to sustain government accountability (Hamilton and Turner, 2009). Such tools are said to make it cheaper for newsrooms to get involved in in-depth investigation (Cohen et al, 2011a). And finally, the third way is by increasing citizens' political participation through their own production and analysis of data (Daniel and Flew, 2010; Cohen et al, 2011b).

How does the integration of programmers in newsrooms - armed with their skills and technical artifacts - really affect the way journalism can contribute to the public good? Although the cultural significance of databases in our societies has been explored (Manovitch, 2001), there has been little empirical investigation into how such socio-technical objects shape knowledge outside of scientific worlds (Bowker, 2006; Hine, 2008). Moreover, while the socio-political implications of technological innovation in news organizations have recently been a major concern for scholars, some of them suggest that emerging work practices in online newsrooms conflict with a strong democratic role of the media in contemporary societies. The fragmentation of journalistic work (Quandt, 2008), the constraints of constant publication, as well as the dramatic increase of imitation between online journalists seem to weaken the watchdog approach to news making (Boczkowski, 2010).

In this article, following a tradition that considers journalism as a form of knowledge (Park, 1940), we analyze how the contemporary integration of programmers in newsrooms challenges established epistemologies of how data can support

investigative journalism. The term “epistemology” is used here with reference to journalists’ knowledge claims about the empirical world. Our intention is not to say whether those claims are valid or not, but rather to identify the norms that make them justifiable beliefs for journalists (Ettema and Glasser, 1987). As epistemological concerns are strongly connected to moral and political concerns, we take the journalistic claim of contributing to the public good or collective justice seriously (Glasser and Ettema, 1989; Lemieux, 2000). And since these epistemologies are strongly connected to technical artifacts promoted by individuals from the computing world, we base our analysis on the way in which Actor-Network Theory envisions technological change (Akrich and Latour, 1992). In particular, we consider that a new technology is introduced by some actors who carry proposed definitions of how it is to be used, and that technical artifacts have embedded definitions and strategies that were not initially designed for news organizations (Schmitz Weiss and Domingo, 2010).

Our claim in this article is that the established epistemology conveyed by the computer-assisted reporting tradition in the United States - relying on the assumption that data can help journalists to set the political agenda through the disclosure of public issues - has been partly challenged by the integration of programmers who mostly come from outside the journalism community. Although they have built on previous achievements by computer-assisted reporters, these programmers and their technical artifacts have conveyed new and challenging epistemological propositions related to

their involvement in open source communities and open government advocacy.

Methods

This study draws on a combination of qualitative research interviews and quantitative news content analyses. The research took place in Chicago, Illinois, in September 2010. We conducted in-depth interviews with fifteen individuals involved in various social worlds. First, we interviewed seven *Chicago Tribune* journalists: two programmer-journalists from the "news applications team"; four investigative reporters involved in database activities; and the *Tribune*'s former vice-president in charge of strategy and development (2000-2008). Second, we interviewed a founder of http:///www.EveryBlock.com, an independent data-driven news website originally launched in Chicago. Third, we interviewed five individuals highly committed to open government advocacy in Chicago. Fourth, we interviewed a couple of Medill School of Journalism professors deeply involved in connecting programmers and journalism. We furthermore quantitatively analyzed two corpuses of data-driven investigative projects: the first one was collected in the printed edition of *The Chicago Tribune* between 2002 and 2009; the second was taken from http://www.ChicagoTribune.com between April 2009 and March 2011. Last, we analyzed many online publications by programmer-journalists.

Using data to disclose public issues: the CAR legacy

Although computer databases were used occasionally in support of investigative

reporting from the late 1960s, it was only in the 1990s that the use of data in investigative reporting became more common in American newspapers (Garrison, 1998a). Despite the wide diversity of practices, this use of data is partially based on the interpretation, supported by specific individuals and groups, that databases are an efficient means to bring public issues to the public's attention and thus to influence the political agenda.

Computer-Assisted Reporting in American journalism

At the end of the 1960s, Philip Meyer was one of the first American journalists to have used computers to do investigative journalism. As he put it in his well-known book, *Precision Journalism: A Reporter's Introduction to Social Science Methods*, first published in 1973: "we journalists would be wrong less often if we adapted to our own use some of the research tools of the social scientists" (Meyer, 1973: 3). According to him, both computers and statistics were effective tools to perform traditional journalistic tasks: "finding facts, inferring causes, pointing to ways to correct social problems, and evaluating the efforts of such corrections" (Meyer, 1973: 4). As a translator between social science and journalism, Meyer's intention was to find better tools to apply the norm of objectivity in American journalism (Schudson, 2001). For this purpose, databases were considered a key technology for journalists as society became more complex and information more abundant. As Meyer put it, "a journalist has to be a database manager, a data processor, and a data analyst" (Meyer, 1991: 1). Initially, a

few pioneers analyzed the databases that they built themselves from data they had collected. In the 1980s it was however still more common to see journalists analyzing existing databases released by government authorities or built from publicly available data (DeFleur, 1997; Cox, 2000).

Since the late 1960s and the pioneer stories published by P. Meyer, C. Jones and D. Burnham (Cox, 2000), the expression "computer-assisted reporting" (CAR) has sometimes been used as a label for stories revealing injustice in society by pointing out the existence and the causes of a social issue, and identifying solutions to it. In 1972, Burnham from the *New York Times* analyzed crime reports and arrest statistics from the New York City Police department, revealing discrepancies among the rates of crimes reported in the city and the arrests made in the precincts (DeFleur, 1997). He pointed out in 1973 that a black person in New York City was forty times more likely to be murdered than a white person. Here, computer data and statistics were used to reveal issues concerning the public good. Another case among many others is Bill Dedman, from The Atlanta Journal-Constitution, who won the Pulitzer Prize in 1989. Combining data from the US Census Bureau and the Federal Financial Institution Examination Council, he revealed racist policies in lending in Atlanta-area financial institutions. In each case, the output was a counter-intuitive piece of information that went against received ideas or ordinary prejudices, and often led to political reforms.

From the mid-1990s, the number of database-oriented projects increased substantially,

spreading to medium-sized and even small news organizations (Garrison, 1998a). With the growing availability of computerized data - on tapes, CD-ROMs, and then online - and the rise of the World Wide Web, the range of journalistic practices involving databases has grown considerably. Computer-assisted reporters have not only used databases to collect, store, organize and retrieve information, but sometimes also to present it in printed supplements or online presentations giving access to education or crime figures (Garrison, 1998b: 261-266). Some of them have even occasionally appealed to citizens to feed their databases in order to improve the reporting (Reisner, 1995). Despite the growing diversity of practices from the mid-1990s, some standards have been collectively set to make databases more regular elements of news-making, e.g. the idea that database-oriented operations are valuable only when they are subordinated to a story idea (Garrison, 1998b: 281); that journalistic norms remain valid in such operations (e.g. checking data for accuracy); and that statistics are crucial to extract the story from the data. The National Institute for Computer-Assisted Reporting (NICAR) - created in 1989 - as well as the handbooks published on the topic (Houston, 1996; DeFleur, 1997; Garrison, 1998b; Houston, Bruzzese and Weinberg, 2002) have been instrumental in exploring and disseminating these standards.

Computer-assisted reporting at The Tribune

Given the diversity of CAR practices since the mid-1990s, it would be difficult to isolate one consistent epistemological model. This is why we analyzed 69 data-driven

stories published in the printed edition of *The Chicago Tribune* between June 2002 and April 2009. One of their co-authors was Darnell Little, who was the *Tribune*'s main computer-assisted reporter at that time. With a background in electrical engineering, Darnell began his career as a software developer for AT&T Bell Labs between 1989 and 1992. A few years later, he left the company and obtained a Master's degree in journalism from Northwestern University. In the following years, he worked as an online reporter at the *Tribune* and then as a computer-assisted reporter. He was also formerly involved in NICAR activities and had many contacts among social scientists.

Each story Darnell was involved in as a computer-assisted reporter rests on data processing or data analysis. Their titles very often emphasize the disclosure of a local issue: "Recession toll worst for whites" (September, 2002), "School spending disparity revealed" (March, 2004), "Drug arrests reveal racial gap" (July, 2007), "Struggling Asians go unnoticed" (March, 2008). Often published on the front page, stories mostly reveal issues and open them to public debate by pointing out injustice, discrimination, and unfair treatment which individuals suffer due to low incomes, gender, race or residence. More or less implicitly, half of all stories call for public intervention or criticize the lack thereof.

These journalistic outputs combine literary accounts and pictures with tables, charts and maps. They mostly deal with local matters of the City of Chicago or of the State of Illinois. Among the topics covered, education comes first (47.8%) and

demographic matters second (20.3%), followed by finances, business and working conditions (8.7%), urban services and transportation (8.7%), crime (7.2%) and officials (4.3%) (Table 1). <TABLE 1>

Stories strongly depend on publicly available data from federal or state authorities. Only a few stories rely on data from authorities related to the City of Chicago (Table 2). As governments release data in computerized form, the computer-assisted reporter seldom designs a new database (only for 7.3% of stories). Moreover, most published articles are based on data released by public officials (60.9%) (Table 3). So most stories are driven by the releasing of data, and the reporter identifies paths in those data, based on statistics. <TABLE 2>

Stories rely strongly on a statistical process developed inside the newsroom (87%). This process is embodied in charts, tables, maps and literary accounts of statistical trends (mean, deciles, etc.) so that readers can easily understand trends. Apart from figures and visual presentation of data, almost every article quotes individual witnesses on a specific issue. Thus, objectivity of figures and statistical analysis is strongly counterbalanced in the storytelling by individual testimonies and descriptions. Moreover, figures and trends are often explained by experts and social scientists who are enlisted to account for them and to put them into perspective (68.1%). Quite often, public officials in charge of this specific issue are asked for their reaction to the data-driven revelation. <TABLE 3>

Most stories reveal public issues on the basis of data analysis, drawing the attention of public opinion and sometimes calling for public intervention. A first category of stories points out local realities based on the data and that should be considered as public issues. The argumentation mainly focuses on showing discrimination between social groups. Given their race, gender, income or location, individuals are likely to be treated in a certain way or to be prevented from obtaining a specific resource - e.g. a black child is forty times more likely than a white child to be in the worst of the worst schools of Chicago ("Still separate, unequal", September, 2004). By forcing public authorities to recognize the issue and to do something about it, such revelations from statistical analysis are often seen as increasing the journalistic contribution to the public good. A second category of stories investigates how a law is enforced locally. Here, statistical analysis shows the unexpected effects of a new reform or the discriminations some groups suffer from in law enforcement, e.g. most minority pupils are not taken into account in school evaluations ("Test scores don't count the neediest", December, 2004). A third category of articles focuses on officials' wrongdoings, criticizing decisions made by some officials and showing that they have been influenced by private interests ("How cash, clout transform Chicago neighborhood", January, 2008).

A consistent epistemological model

A consistent epistemological model appears throughout these stories, although this

model is not representative of every CAR initiative in the United States. In this model, public data released by Federal and State governments are statistically processed to reveal local issues and to influence the political agenda. The underlying assumption is that computer data combined with statistics can serve to reveal issues that citizens and journalists cannot fully embrace from their own individual perspective.

In this model, data have no journalistic value on their own. The reporter has to find the hidden "story" in the data:

> "The story is coming from the data and I'm really the only person who can really look and analyze the data. So I've got to go in there, look at it, whether it was census, education, statistics, and basically say, okay, I think this is the story we want to do from this data." (Darnell Little, September 10, 2010)

In this model, data must allow for sampling. Based on social science, controlled sampling produces intelligibility for the whole city or State. They also have to contain information about social groups: sex, race, incomes or location are major variables. As Darnell explains, it is often hard to identify social groups in data:

> "Most states don't disaggregate education data by ethnicity. So if someone's Asian, they just classify them as Asian. (...) But we were able to dig enough to get at least enough information to also show that in public schools you see the same thing. Southeast Asians, Cambodians, Thais, Mongs, even to a degree Vietnamese, struggle a lot more at schools than other Asian ethnicities. "

(Darnell Little, September 10, 2010)

Reporters believe that since data are produced by officials, their categorization might be manipulated. This is why they have long been reluctant to use data from city officials - especially the Chicago Police department which they have generally suspected of manipulating data.

Although online access to data has often been given to *Tribune* readers since 1996 (especially for test scores or crimes), Darnell consider that it is his duty to point out trends and to give explanations to the public through literary accounts and graphs:

> "I think we just can't make raw data available by itself and then expect all of that other stuff that happens just where it's automatic. If we had enough raw data available on the web, there won't be as much of a need for investigative reporting, because people will be able to connect all these dots. That's not true."
>
> (Darnell Little, September 10, 2010)

This epistemological model is partly a continuation of a traditional conception of how journalism can contribute to democracy - especially the muckrakers' tradition and its reformist approach to social change (Feldstein, 2006). By revealing truths that can be found in publicly available data, this approach intends to inform public debate and influence the political agenda. Several investigative journalists still share this approach at the *Tribune*. Yet the rise of so-called "programmer-journalists", notably in Chicago, in the mid-2000s, has partially challenged this model.

New epistemological propositions made by "programmer-journalists"

Since the mid-2000s, newcomers have joined the newsrooms of some established newspapers in the United States and Great Britain, as well as independent news organizations and start-ups specialized in data-driven journalism. Coming mostly from the computer worlds, these recruits often claim to be "programmer-journalists" although there is no consensus among them about this title (Pilhofer, 2010). In the newsroom, their main task is to design "news applications" - online presentations, databases, interactive maps, etc. - involving a wide range of computer skills to collect, process, analyze and visualize data sets. Because of their involvement in open source communities and open government advocacy, these programmers and their technical artifacts have conveyed new epistemological propositions challenging the established model of how data can support investigative journalism. By analyzing these propositions, we do not mean to suggest that the link between databases and journalism is new or unique to programmer-journalists. Rather, we seek to show how, building on previous historical developments, their propositions challenge long-standing journalistic epistemological principles in a certain way.

New connections between journalists and programmers

In terms of social worlds, two dynamics account for the new connections between journalists and programmers in the United States in the mid-2000s. First, a handful of journalists have been hired in several newspapers since 2005 to design web applications

that have a news purpose. Adrian Holovaty at the *Washington Post*, Aron Pilhofer at the *New York Times*, Matt Waite at the *St-Petersburg Times* or Ben Welsh at the *Los Angeles Times* perfectly embody that shift. They all share the concern that newspapers should set up dedicated units staffed with people familiar with journalism as much as code. Pilhofer was one of the first to achieve this when he convinced *Times*' editors to create the "Interactive news technology department" (Royal, 2010). Most of these journalists have a background in CAR and have acquired sound skills in programming languages (Python, Ruby, PHP, etc.). Some of them are also committed in open software communities, especially Adrian Holovaty, who has been highly regarded in open source communities as the co-creator of Django, a web framework originally designed to make it easier for programmers to meet the intensive deadlines of a newsroom.

The second dynamic corresponds to a large number of data-driven initiatives that have emerged under the banner of open government advocacy since 2005. Web entrepreneurs, activists, civic hackers, computer enthusiasts and journalists have progressively joined forces to strengthen government accountability and citizen participation through the release of government data (Lathrop and Ruma, 2010). First in major American cities, some of them tried to promote the release of public data by municipalities. In this respect, Washington DC has been a reference with its data portal launched in 2004. Sharing the normative beliefs about freedom of information promoted

by the open software movement (Holtgrewe and Werle, 2001), they have designed websites and online applications resting on city government data that give citizens the opportunity to know what is going on in their city (as regards crime, transportation, etc.) and to control their elected officials. The movement has spread to the federal level, and many data-driven online applications have been designed based on the same political concerns. Launched in 2007, http://MapLight.org for instance was designed by Daniel Newman - a programmer, entrepreneur and political activist - to correlate lawmakers' voting records with the money they have received from special-interest groups (Lathrop and Ruma, 2010: 223-232).

Most of these projects were originally designed outside the journalistic world, but open government advocacy has paved the way for new connections between journalists and programmers. Many individuals from the computer world have been attracted by the underlying political concern that computer skills can be devoted to the public good. Working for news organizations has thus become rewarding for programmers although financial opportunities are often considerably lower than in banks or advertising companies. Brian Boyer - a former web programmer who in 2009 set up a "news applications team" within the *Chicago Tribune*'s newsroom - stresses this:

> "Journalism is sort of democratically enabling things to get better. It's about informing people, voting, things like that, and democracy." (Brian Boyer,

September 9, 2010)

In Chicago, a scene has emerged where programmers working for private firms, computer enthusiasts involved in open software communities and activists committed to open government advocacy have established contact. Joe Germuska perfectly embodies this: when this programmer involved in open software left his previous job for a commercial company, he began to attend open government meetings more regularly. That was how he met Brian Boyer and Adrian Holovaty, and finally got hired as a programmer-journalist at the *Tribune* in 2009:

> "Well, the open government project started during that same period when I met Brian Boyer. It was all part of what I started doing when I left my last job. About four months before I finished the other job, I went to a conference, which is actually, where I met Dan O'Neil and Adrian Holovaty and some of the other guys" (Joe Germuska, September 14, 2010)

With the support of foundations - the Sunlight Foundation and the Knight Foundation, which has largely funded news projects carried by programmers (Lewis, 2011) - and universities offering journalistic trainings for programmers - Medill School of Journalism -, a handful of individuals from the computer worlds have entered the journalistic world in Chicago. Most of them do not share the statistical culture of computer-assisted reporters and have no connection with social science. Relying on some normative principles shared by open source communities and open government

advocacy, they have brought in epistemological propositions of how data can support investigative journalism. Although they build partly on previous developments reflected in the CAR tradition, these propositions challenge some established journalistic conceptions. From the study of local online news projects initiated in Chicago since the mid-2000s, and interviews with some of their founders and staff members, we have identified three major propositions.

Proposition 1: "News is structured information"

Their first claim is that news in itself should be viewed as computer processable data, and not only as a story hidden in the data. This implies not only that the online publication of a database should become a legitimate dimension of news-making, but also, more profoundly, that each task which journalists perform in the news-making process - collecting, analyzing and presenting information - could be more effectively performed with computers since this information is basically "structured". In Chicago, Adrian Holovaty has been one of the major advocates of this claim:

> "Much of what local journalists collect day-to-day is structured information: the type of information that can be sliced-and-diced, in an automated fashion, by computers. (...) For example, say a newspaper has written a story about a local fire. (...) what I really want to be able to do is explore the raw facts of that story, one by one, with layers of attribution, and an infrastructure for comparing the details of the fire - date, time, place, victims, fire station number, distance from

> fire department, names and years experience of firemen on the scene, time it took for firemen to arrive - with the details of previous fires. And subsequent fires, whenever they happen." (Adrian Holovaty, 2006)

Considering news as structured information has several implications. The first is that journalism and programming should be viewed as inseparable in the newsroom because they both involve dealing with databases and structured information. Although it is partly a continuation of several CAR initiatives in the 1990s, this statement extends the scope of computer treatment to any event reporters have to deal with. In that regard, the growing availability of data sets released by local governments from the mid-2000s has made this statement more realistic than it could have been before.

A second implication is that new knowledge can be produced from the combination of heterogeneous data sets. This claim was already prominent in the CAR tradition from the early days, but the difference lies in Holovaty's expectations concerning programming techniques. In his opinion, these techniques allow for databases to be designed from any heterogeneous data – whether it be numbers, words, pictures, etc. - and to be automatically combined so that the potential link between data sets can be made visible through online presentations. This statement is strongly connected to huge expectations regarding computer techniques (mashups and web applications) that are seen as making data transparent to the audience.

In Chicago, several news projects have embodied that claim. The most radical

one was designed by Holovaty in 2005 when he built a mashup over the Chicago police department's website and Google Map. This former website, called ChicagoCrimes, allowed one to check the crimes that occurred in a particular neighborhood, and thus to search the crime database on the block level. It also allowed for various cross tabulations, by crime category, area type, street name, zip code and others. In 2007 Holovaty launched http://www.EveryBlock.com, which has been supported by the Knight Foundation. Extending the scope of ChicagoCrimes by adding new data, this website enables one, on the block level, to check not only crimes but also a vast range of public data from public records (building permits, liquor license applications) or private records (real estate listings, school reviews). As Daniel O'Neil, one of the co-founders of EveryBlock explains, the combination of heterogeneous data sets may generate useful knowledge:

> "Okay, let's see. You got a liquor license and then crime went up. I think that it's fair to mention that. I think it's a reasonable piece of data - not even data - insight to use in planning, right?" (Daniel O'Neil, September 8, 2010)

Considering news as structured information implies a significant break with the first epistemological model of how data can support investigative journalism. Most of the programmer-journalists we interviewed do not share the belief that public issues are hidden in the data and that their primary role would be to reveal such issues using statistical methods. Most of them - especially at EveryBlock's - do not consider

statistics as a major tool because, in their opinion, data do not hide anything if they are granular and complete. In fact, Holovaty and his staff view granularity and completeness, closely related to open government advocacy, as a major issue for journalism:

> "Aggregate is a way to lie obviously. So data has to be granular, that is to say incident level" (Daniel O'Neil, September 8, 2010)

Our aim here is not to question whether this approach to what makes the value of data sets is valid or not, but rather to point out the epistemological assumptions made by these programmer-journalists. In that regard, as they put the emphasis on data granularity and completeness, they depart from the idea that intelligibility is the result of statistical techniques, and especially sampling. The idea that one can produce new knowledge from the analysis of a sample is not familiar to them. Instead, they believe that intelligibility is the result of affording access - thanks to computer tools - to complete and granular data from which citizens are usually kept away.

Proposition 2: "Designing research tools for the audience"

Because of their involvement in open software communities and open government advocacy, programmer-journalists view public access to information - whether it is code or public information - as a major issue. This is why they all put the emphasis on giving the audience the largest and easiest access to the data. They claim that readers should be given the opportunity not only to check the data but also to combine them and to use

them for other goals. Although giving access to data has been an occasional concern among online journalists since the 1990's, this proposition actually challenges established conventions of how data can support watchdog journalism.

To users, such websites generally appear to be quite sober: they simply have to write down their address or zip code and they have access to a map and several lists that they can browse. But from the designers' point of view, such news projects offer readers a convenient "decision-making tool" or "research tool" in their daily life through simple and standardized access to data:

> "It could be a citizen who finds out about something and goes and looks at it, and then makes a decision about the neighborhood, about what they're going to bring back to their block club or what - you know, if I'm going to live here or not. So it's a decision-making tool and a research tool for everybody." (Daniel O'Neil, September 8, 2010)

At EveryBlock, the main concern is to give readers tools to help them solve their problems as consumers, citizens and public services users:

> In Chicago, we have this thing called Community Area Policing. And they use our data, the police's data filtered through us, in those kind of ways, where they seek to make connections between a business and an increase in crime. And there's all sorts of that kind of stuff that I think that frankly there's not enough of at this time. (Daniel O'Neil, September 8, 2010)

At the *Tribune*, Brian Boyer and his team also put the emphasis on giving out the data - through friendly applications - so that the readers can do their own research on how an issue affects their personal situation. One example is the application they designed in 2009, that enables users to find, for any nursing home in the Chicago area, data on how many residents are mentally ill, how many of them are felons or registered sex offenders, and how many hours nurses spend with each resident per day. Offering such a research tool to the audience is seen as a way to personalize stories:

> "I think that this application told a very specific story to a lot of people, but it was about telling a specific story to specific people. It's all about knowing your audience and telling something specific instead of broadly generically saying it." (Brian Boyer, September 9, 2010)

These programmer-journalists, much more widely than their predecessors, argue that people are able to extract the meanings of data on their own, and eventually to make their own moral claim. Although there are differences between the two news organizations, this proposition seriously challenges established journalistic conventions. Not only because American journalists have considered that the growing complexity of the world has made it necessary for them to give interpretations and analyses to their audiences (Schudson, 1978), but more specifically because it challenges the epistemological model conveyed by computer-assisted reporters. According to this proposition, journalists should not centralize the construction of a moral claim with the

support of data analysis so much: readers are considered here as legitimate and active contributors in this process.

Proposition 3: "Reduce the dependence on government agendas"

Transparency is a major concern among open government activists, although the political implications of that norm have been questioned both outside (Birchall, 2011) and inside the movement (Lathrop and Ruma, 2010: 267-72). Yet the basic assumption is that the release of data by public authorities and the design of data-driven applications can strengthen government accountability and communication between the different levels of government (Lathrop and Ruma, 2010). Sharing this general concern, programmer-journalists view it as a challenge to the established ways in which journalists use data in support of government accountability.

EveryBlock's staff expects users to control their elected officials through the manipulation of data. For instance, delivering data concerning city services allows citizens to discover how internal political conflicts may impact the quality of services they get:

> "From the citizen point of view honestly I had as a goal to allow people to see inequities, to see if there is any punishment. If an alderman from a particular area gets up in city council and lambaste the mayor. Which of course never happens, but if it did, what services would be held back. I am personally interested in that subject." (Daniel O'Neil, September 8, 2010)

This is of particular concern to O'Neil, who launched another website in 2009, called http://www.CityPayments.org. The site allows users to check payments and contracts made by the City of Chicago with vendors, to comment on them, and to label them as "goofy". At the *Tribune*, similar projects have been designed that allow readers, for instance, to look through the spending of their alderman:

> "A reader found that one of the aldermen had been renting a property from himself, for his family. And so he emailed us, and we emailed Hal Dardick, and said, "Hal, there may be some more stuff in this data." (...) The first day the story was "Alderman doing unethical things." The second day's story was, "Alderman doing potentially criminal things." In the end, nothing really happened of it, but I like to think that putting all the data online, that's the sort of thing we're going to continue to enable." (Brian Boyer, September 9, 2010)

But to what extent would such projects, aimed at making officials' wrongdoings visible, challenge established journalistic conventions? Before the arrival of programmer-journalists in the *Tribune*'s newsroom, data were regularly processed to reveal wrongdoings. Yet this proposition is challenging for two main reasons. The first is that such projects rely more on data sets from city governments. Taking advantage of the growing availability of city-level data, programmer-journalists view such data as an opportunity to increase control over city officials. The second reason is more fundamental. According to these programmer-journalists, the availability of both

sustained data sets and database techniques makes it possible for journalists to be less dependent on the release of data by governments - even if most data are produced by public authorities. As databases are regularly and automatically updated, online applications allow journalists and ordinary citizens to check decisions made by city officials and to monitor possible wrongdoings. As a consequence, the revelation of wrongdoings would be less dependent on strategic moves made by sources within the administration. We analyzed 37 database projects published on ChicagoTribune.com between April 2009 and March 2011. Our analysis clearly indicates a drop in the percentage of projects that follow the release of data by governments (Table 4).

<TABLE 4>

Strongly related to their involvement in open source communities and open government advocacy, programmer-journalists in Chicago have conveyed three propositions that partially challenge the first epistemological model supported by the CAR tradition. Rather than revealing truths that can be found in publicly available data with the support of statistics, they suggest another approach where truths would be brought out through the access, combination and processing of data from which citizens are usually kept away. We shall now discuss the scope of such claims that have been controversial among American journalists.

Discussion and controversies

Recent connections between journalism and the computer worlds have partly renewed

the epistemological basis of how data-driven journalism may contribute to the public good. With the support of various institutions, these challenging claims have found their way into some of the most established news organizations. Although the specific context of our study does not enable us to draw more general conclusions about data-driven journalism in the United States or abroad, we shall point out the main controversial issues among journalists that might affect their real impact on newswork practices.

Embedding political considerations in technology

When asked about the changes taking place within their profession, journalists often refer to technology as a self-sufficient explanatory factor (Örnebring, 2010). And the integration of programmers within the newsroom may be viewed as another example of technology-driven innovation. But even though these individuals are collectively deeply committed to designing technical artifacts, their commitment is strongly connected with normative considerations. This is especially true in Chicago, where the "hacker community" is often depicted as liberal and progressive. Indeed, the programmers we interviewed share the general idea that technical artifacts (web applications, web frameworks and programming languages) hold great promise both for news organizations and democracy. Because of their involvement in social movements such as open software and open government advocacies, they have promoted specific points of view as to how data-driven artifacts should be used in news organizations (Schmitz

Weiss and Domingo, 2010). This type of embeddedness of political considerations in technical artifacts has often been analyzed by scholars (Winner, 1986), especially for the Internet (Turner, 2006).

Considering society as an engineering system, programmer-journalists share this idea that such collective entities as "social groups" or "government" should not be taken for granted. Technology here is used to deconstruct these constituted social entities. For instance, governments are deconstructed as social entities so that any reader can monitor the effects that local elected officials have on their personal environment. As we explained, this represents a significant departure from the epistemological model supported by computer-assisted reporters. In this approach, social entities are very important, as the cornerstone of collective justice.

Within news organizations this approach to technology, as well as the three epistemological claims, have been questioned intensely, mainly in two ways.

What makes the journalistic value of data?

Among journalists, there has been a controversy regarding the conditions under which data sets may have a real journalistic value. As described above, programmers share a strong belief in data transparency. According to them, the data cannot lie or hide anything if they are granular, complete and regularly updated. At the *Tribune*, some journalists have questioned this idea of data transparency, claiming that, on their own, the data cannot be meaningful to most people. Jason Grotto - a *Tribune* investigative

reporter who used to be deeply involved in NICAR - shares this opinion:

> "I question the usefulness of it because data by itself, what does it tell you? Okay, you had some crimes in your neighborhood, but what does it mean? Journalists interpret, filter, do all kinds of things that I think are important. That's what we bring to bear. So data in and of itself isn't necessarily valuable." (Jason Grotto, September 16, 2010)

This approach has also been criticized for underestimating data manipulation by political authorities. For instance, in April 2009, Los Angeles Times' journalists found that EveryBlock has released faulty government data (Welsh, 2009). At the *Tribune*, editors decided not to release crime data automatically supplied by the Chicago Police Department because they suspected it of manipulating the data. Instead, a reporter was asked to collect and check data from two different sources - the police and medical examiners - in order to produce an accurate 'homicide map' in Chicago. This controversy is not over, but the general belief in data transparency has generated much resistance among journalists.

How far can readers participate in the identification of a public issue?

Another controversial issue among journalists has been readers' participation in the identification of new public issues, from data analysis. As explained above, designing research tools for readers is a major concern for programmer-journalists. Many journalists see this as expecting too much from ordinary readers, most of whom do not

have the necessary skills to sort and interpret data on their own. At the *Tribune*, programmer-journalists have kept their distance from the radical stance of Adrian Holovaty:

> "I think those guys at EveryBlock's are doing some exciting work, but I don't use it very much. I haven't really figured out how to fit it into my routine. And so we didn't want it to be just another blog data application. And so we've worked a lot with the politics reporters to try and get them to provide narrative and back story and context." (Joe Germuska, September 14, 2010)

As a result, this established news organization generally does not expect as much from its readers. Data-driven applications are considered to offer specific stories for specific readers, but *Tribune* journalists still consider their professional role to be the identification of public issues. Indeed, there are some limits affecting readers' participation in the identification of a public issue. It is not always a matter of common sense to identify whether an alderman's expense is legally and morally justified or not, and journalists are still major actors in this process. As the *Tribune*'s former vice-president in charge of development explained, taking the example of school report cards, there will always be a difference between the readers' and reporters' approach to data:

> "A user wants to know, "How good is my school this year? Did it get better than it was last year?" And a reporter will say, "Boy, that school got a lot better and

there's no evidence why it should. I need to figure that out." Well, you need both. You want to satisfy the needs of the people who content questions that they want answered. But there's always things that they'll never think about that the journalists should be trained to ferret out and communicate." (Owen Youngman, September 16, 2010)

It would be hasty to draw general conclusions about such trends, but controversies both inside and outside news organizations suggest that some of the original claims made by programmers might now be aligned with more traditional journalistic professional standards.

Conclusion

This paper has analyzed how the rise of programmer-journalists in newsrooms challenges established epistemologies of how data can support investigative journalism. Although they have built on previous historical developments, these new forms of data-driven journalism rely on epistemological considerations that are strongly connected to social movements within the computer world. Rather than revealing truths that can be hidden in publicly available data with the support of statistics, programmer-journalists propose another approach where truths are disclosed through the accessing, combination and processing of complete data. These transforming epistemologies raise the question of how contemporary journalism contributes to the public good by making things visible and pointing out new issues.

This inquiry highlights the importance of epistemological and normative considerations in the growing articulation between journalists and programmers. As Schudson (2010) has put it, contemporary uses of databases in news have developed thanks to a more general democratic process. This is why research on contemporary data-driven journalism should investigate both the epistemological and the socio-political meanings that are collectively assigned and discussed at the interface between these worlds.

The empirical context of this study is restricted to the city of Chicago, which is particular when it comes to computer worlds and news organizations. Yet the rise of programmer-journalists is a much wider phenomenon that concerns many news organizations in the western world (Gray, Chambers and Bounegru, 2012). Future research should therefore compare practices between countries, taking into account the variations in hacker and journalistic cultures across countries. For instance, the fact that CAR is poorly known in some European countries (e.g. France) may have some effects on how data-driven journalism is actually put into practice.

Finally, this inquiry is based on a set of methods which emphasize the articulation between social worlds. Further research should investigate single organizations to analyze how these programmer-journalists actually integrate the news-making process. Ethnography could offer an appropriate way to study how the news-making process and professional identities are both currently transformed.

Acknowledgements

This work was supported by the Agence Nationale de la Recherche and the French Ministry of Culture.

This article has greatly benefited from close readings by Ashveen Peerbaye, Patrice Flichy, Dominique Cardon, Nicolas Auray and Ignacio Siles. We also thank the editors and two anonymous reviewers for their most helpful suggestions.

References

Akrich M and Latour B (1992) A summary of a convenient vocabulary for the semiotics of human and non-human assemblies. In: Bijker W, Law J (eds) *Shaping Technology-building Society: Studies in Sociotechnical Change*. Cambridge, MA: MIT Press, 205-24.

Arthur C (2010) Analysing data is the future for journalists, says Tim Berners-Lee. *The Guardian*, November 22.

Birchall C (2011) Introduction to 'secrecy and transparency': the politics of opacity and openness. *Theory Culture Society* 28(7-8): 7-25.

Boczkowski PJ (2010) *News at work: Imitation in an age of information abundance.* Chicago, Illinois: University of Chicago Press.

Bowker G (2006) *Memory practices in the sciences*. Cambridge: MIT Press.

Cohen S, Hamilton J and Turner F (2011a) Computational journalism. *Communications of the ACM* 54(10): 66-71.

Cohen S, Li C, Yang J and Yu C (2011b) Computational journalism: A call to arms to database researchers. 5th biennial conference on innovative data systems research, January 9-12, 2011, Asilomar, California, USA.

Cox M (2000) The development of computer-assisted reporting. A paper presented to the newspaper division, association for education in journalism and mass

communication, southeast colloquium, University of North Carolina, Chapel Hill.

Daniel A and Flew T (2010) The Guardian reportage of the UK MP expenses scandal: a case study of computational journalism. Paper presented at communications policy and research forum, Sydney.

DeFleur M (1997) *Computer-assisted investigative reporting: development and methodology*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ettema JS and Glasser TL (1987) On the Epistemology of investigative Journalism. In: Gurevitch M and Levy MR (eds) *Mass Communication Review Yearbook*, 6. London: Sage.

Feldstein M (2006) A Muckraking model. Investigative reporting cycles in American history. *The Harvard International Journal of Press/Politics* 11(2): 105-120.

Garrison B (1998a) Newspaper size as a factor in use of computer-assisted reporting. Paper presented to the Communication Technology and Policy Division of the Association for Education in Journalism and Mass Communication, Baltimore.

Garrison B (1998b) *Computer-assisted reporting* (2nd edition). Mahwah: Lawrence Erlbaum Associates.

Glasser TL and Ettema JS (1989) Investigative journalism and the moral order. *Critical Studies in Mass Communication* 6(1): 1-20.

Gray J, Chambers L and Bounegru L (2012) *The Data Journalism Handbook*. O'Reilly Media.

Hamilton JT and Turner F (2009) Accountability through algorithm: Developing the field of computational journalism. Center for advanced study in the behavioral sciences summer workshop.

Hine C (2008) *Systematics as cyberscience: computers, change and continuity in Science*. MIT Press.

Holovaty A (2006) A fundamental way newspaper sites need to change. Available at: http://www.holovaty.com/writing/fundamental-change/

Holtgrewe U and Werle R (2001) De-commodifying software? Open source software between business strategy and social movement. *Science Studies* 14(2): 43-65.

Houston B (1996) *Computer-assisted reporting: a practical guide*. New York: St. Martin's Press.

Houston B, Bruzzese L and Weinberg S (2002) *The investigative reporter's handbook. A guide to documents, databases and techniques*. New York: Bedford/St. Martin's Press.

Lathrop D and Ruma R (2010) *Open government: collaboration, transparency and participation in practice*. Sebastopol, CA: O'Reilly Media.

Lemieux C (2000) *Mauvaise presse: une sociologie compréhensive du travail journalistique et de ses critiques*. Paris: Métailié.

Lewis S (2011) Journalism innovation and participation: An analysis of the Knight News Challenge. *International Journal of Communication* 5: 1623-48.

Manovitch L (2001) *The Language of New Media*. Cambridge, MA: MIT Press.

Meyer P (1973) *Precision journalism: A reporter's introduction to social science methods*. Bloomington & Indianapolis: Indiana University Press.

Meyer P (1991) *The new precision journalism*. Bloomington & Indianapolis: Indiana University Press.

Örnebring H (2010) Technology and journalism-as-labour: historical perspectives. *Journalism* 11(1): 57-74.

Park RE (1940) News as a form of knowledge: A chapter in the sociology of knowledge. *The American Journal of Sociology* 45(5): 669-86.

Pilhofer (2010) Programmer-Journalist? Hacker-Journalist? Our identity crisis. Mediashift IdeaLab, available at: www.pbs.org/idealab/2010/04/programmer-journalist-hacker-journalist-our-identity-crisis107.html

Quandt T (2008) News tuning and content management: An observation study of old and new routines in German online newsrooms. In: Paterson C, Domingo D

(eds), *Making online news: The ethnography of new media production*. New York: Peter Lang, 76-97.

Reisner N (1995) On the beat: Computer-assisted reporting is not just for projects anymore. *American Journalism Review* 17(2): 44-7.

Royal C (2010) The journalist as programmer: A case study of the New York Times interactive news technology department. Paper presented at the International Symposium in Online Journalism, University of Texas, April 2010.

Schmitz Weiss A and Domingo D, 2010 Innovations processes in online newsrooms as actor-networks and community of practice. *New media & society* 12(7): 1156-71.

Schudson M (1978) *Discovering the news: A social history of American newspapers*. New York: Basic Books.

Schudson M (2001) The objectivity norm in American journalism. *Journalism* 2(2): 149-70.

Schudson M (2010) Political observatories, databases & news in the emerging ecology of public information. *Daedalus*, 139: 100-9.

Turner F (2006) *From Counterculture to Cyberculture. Stewart Brand, the Whole Earth Network, and the Rise of Digital Utopianism*. Chicago: University of Chicago Press.

Winner L (1986) *The whale and the reactor*. Chicago: University of Chicago Press.

Table 1. Topics of the *Tribune*'s database stories (2002-2009)

topics n=69	percentage
education	46.4%
demography, incomes & ethnic minorities	20.3%
finances, business & working conditions	8.7%
urban services & transportation	7.2%
crime	5.8%
officials	4.4%
other topics	7.2%

The table shows the distribution of topics among the 69 data driven stories published in the *Tribune*'s printed edition between June 2002 and April 2009 (source: personal archives of Darnell Little).

Table 2. The political scale of data sources (2002-2009)

Political scale of data sources n=69	percentage
Federal administration	39.1%
State of Illinois	47.8%
City of Chicago	13.1%

The table shows the distribution of political scale of data sources among the 69 data driven stories published in the *Tribune*'s printed edition between June 2002 and April 2009 (source: personal archives of Darnell Little).

Table 3. Main features of the *Tribune*'s database stories (2002-2009)

Main features of the Tribune's database stories n=69	percentage
stories following the release of data	60.9%
online access to the database (www.chicagotribune.com)	30.4%
interviews or qualitative research	92.8%
charts, tables or maps	89.9%
quotation of social scientists	68.1%
database designed by *Tribune* staff	7.3%
database statistically analyzed by *Tribune* staff	87%

The table shows the main features of the database stories published in the *Tribune*'s printed edition between June 2002 and April 2009 (source: personal archives of Darnell Little).

Table 4. Main features of the *Tribune*'s database projects (2009-2011)

Main features of the database projects n=37	Percentage
database designed by *Tribune* staff	47.2% (7.3)
database statistically analyzed by *Tribune* staff	50% (87)
stories related to the database project	88.9% (100)
online access to the database	88.9% (30.4)
online access to raw data	69.4% (0)
project following the release of data	25% (60.9)
quotation of experts in related stories	44.4% (68.1)
mainly focused on a local scale (State of Illinois or below)	94.4% (88.4)

Note: 47.2% of database projects published on ChicagoTribune.com between April 2009 and March 2011 were designed by *Tribune* staff, as opposed to 7.3% between 2002 and 2009.