

THE PROMISE OF COMPUTATIONAL JOURNALISM

Terry Flew, Christina Spurgeon, Anna Daniels, Adam Swift

Computational journalism involves the application of software and technologies to the activities of journalism, and it draws from the fields of computer science, the social sciences, and media and communications. New technologies may enhance the traditional aims of journalism, or may initiate greater interaction between journalists and information and communication technology (ICT) specialists. The enhanced use of computing in news production is related in particular to three factors: larger government data sets becoming more widely available; the increasingly sophisticated and ubiquitous nature of software; and the developing digital economy. Drawing upon international examples, this paper argues that computational journalism techniques may provide new foundations for original investigative journalism and increase the scope for new forms of interaction with readers. Computer journalism provides a major opportunity to enhance the production of original investigative journalism, and to attract and retain readers online.

KEYWORDS citizen journalism; computational journalism; data-based reporting; data visualisation; ubiquitous computing;

What is Computational Journalism?

At its simplest, computational journalism involves the application of computing to journalism. This means not simply the application of computing technologies to journalism – journalism has of course long been engaged with information and communication technologies (ICTs) since the modern era – but active engagement with techniques for the large-scale manipulation of data using computing software to enable new ways to access, organise and present information. Hamilton and Turner (2009, 2) define computational journalism as:

the combination of algorithms, data, and knowledge from the social sciences to supplement the accountability function of journalism. In some ways computational journalism builds on two familiar approaches, computer-assisted reporting (CAR) and the use of social science tools in journalism. Like these models, computational journalism aims to enable reporters to explore increasingly large amounts of structured and unstructured information as they search for stories.

The broad definition of computational journalism as the application of computation to the activities of journalism is problematic in that computers have been central to journalism since the moment they replaced typewriters on news desks. A more refined definition of computational journalism would include here computational tools and methods associated with quantitative accountability. Meyer has identified such methods as including statistical analysis, polling, surveying, and the observation, collection, and interpretation of information from public records, interviews, and direct participation (2009, 3). In order to understand what is new about computational journalism, we need to distinguish between *computers as tools* and *computation as theory* (Miller and Page 2007). The capabilities of computers as tools can be readily understood by learning how to use the various software packages that they come with. Learning how to use computational tools is not the same as learning computational techniques, or what Miller and Page describe as “the processes underlying the

computations, as opposed to the computations themselves” (Miller and Page 2007, 77). These include processes such as searching, correlating, filtering, identifying patterns, and so forth. Processes that have historically been undertaken by humans, but that are increasingly being performed with greater speed and accuracy by computational devices.

One implication of extending computational methods and techniques to journalism is that computational journalism can bring information technology specialists and journalists together in order to develop new computing tools that are associated with the aim of providing information that is accurate, original, reliable, and socially useful. To take one example, the process of news gathering corresponds closely to an area of computer science known as sense-making, or the process of generating new knowledge (making sense) out of one’s experiences in the world (Klein, Moon and Hoffman 2006). Humanistic perspectives have related sense making to human cognition, through processes such as creativity, curiosity, comprehension, mental modelling and situated awareness (*c.f.* Dervin *et. al.* 2003): these have been the approaches to sense-making that journalism, and journalism education, has traditionally drawn upon. From the perspective of human-computer interaction, the question of sense making is bound up with the development of intelligent systems, which should be able to:

- Fuse large amounts of data into succinct meanings;
- Process meanings in contextually relative ways;
- Enable humans to achieve insights from this data fusion and processing;
- Infer hypotheses that humans are considering;
- Enable people to have access to the intuitions of others;
- Present information in relevant ways that enhance the tacit knowledge of humans about the subject matter (Klein, Moon and Hoffman 2006).

Our understanding of computational journalism can be extended by exploring not only the various uses journalists make of computational tools, but by examining the underlying principles and processes through which these tools are developed and for which they are employed. Computational journalism is not about getting journalists to think or act like computers, but enabling them to use computing devices to tackle problems beyond the scope of everyday action prior to the age of computing. To this end, computational journalism recognises that computer science’s contribution extends beyond number crunching and search-based activities to offer computational abstractions and methods based more on conceptualisation than programming. Wing (2006) argues that, regardless of whether they are executed by human or machine, the processes of *computational thinking* involve solving problems, designing systems, and understanding intended behaviour – concepts fundamental to computer science. Both computational thinking and the more refined field of *computational intelligence*, as it has been defined and developed in the computer sciences, have much to offer to journalism and are discussed here in areas of data-gathering, information brokering, and knowledge creation as they occur within complex adaptive systems, both social and technological.

Since the development of computing in the mid-1940s machine or artificial intelligence has become a significant sub-domain of computer science. Defined as the “science of making machines to do things that require intelligence if done by a man (sic.)” (Minsky 1968, v), artificial intelligence works from the idea that any commercial or industrial task that could benefit from even a limited degree of intelligence could, in principle, be performed better through computing designed to execute “intelligent” tasks with precision, speed, and reliability that exist beyond normal human capabilities (Boden 1985, 98). Classic subjects of investigation within this field are vision, language and speech processing, robotics, knowledge representation and reasoning, problem solving, machine learning, expert systems, man-machine interaction, and artificial life (Aarts *et. al.* 2006). While some of these

investigations offer little to journalism, real-world applications of machine-based intelligences that exhibit the ability to reason, apply learned knowledge, manipulate information, and to think abstractly do. For example, investigations that require journalists to understand, interpret and relate information may make use of rule-based or expert systems that ‘argue’ and ‘reason’ using data and relations among data, applying heuristic rules to combine and transform data in order to reach valid conclusions or to develop new pieces of information. And investigations that involve the search, retrieval, or cataloguing of specific pieces of information from near unlimited data require intelligent means of support. Likewise, creating and expressing meaningful new representations from large amounts of data requires the intelligent manipulation of variable elements.

Taylor (2010) shows three areas where computational processes can play a fundamental role in humanistic investigation and analysis: automation; algorithms; and abstraction. *Automation* alleviates activities such as data gathering and interpretation, number crunching, network analysis, sorting, and processing that would otherwise need to be done manually; *algorithms* allow operators to follow predefined steps needed to accomplish certain goals, identify problems, find suitable solutions in a large set of alternatives, and verify information in a reliable, consistent and efficient manner; and *abstraction* enables the qualification of different levels or perspectives from which an idea may be presented or new directions that may be explored. Following this, we identify three domains, or areas of investigation, in which the above processes occur. Firstly, there is *data* itself – raw data that is generated through a range of diverse recording, sensory and monitory devices that can be filtered and categorised for further use. It is important to remember that data refers not only to numeric and statistical records, but to any form of information represented in a digitised format including text, audio, photographic, and video files. The application of computational processes and techniques to mine, categorise, filter, amplify and transform this data into discernable trends, patterns and accurate and meaningful summaries generates *information*. However, regardless of whether such processes are undertaken by human or machine, the information that is generated from data cannot be separated from the social relationships and cultural references which give it meaning and value. This becomes even more pertinent when information is represented through the generation of new ideas or new creative expressions. This point is recognised by Brown and Duguid (2000) who, in distinguishing between information and *knowledge*, emphasise how “the importance of people as creators and carriers of knowledge is forcing organisations to realise that knowledge lies less in its databases than in its people” (Brown and Duguid 2000, 121).

The key to new computational approaches, of which computational journalism is one example, is to see human-based and computer-driven approaches to sense making as complementary or supplementary. Computational techniques can improve the quality of investigative reporting and news journalism by greatly decreasing the time spent gathering and processing information from increasingly diverse sources. Computational techniques enable journalists to make more effective use of data by offering not only greater search and retrieval speeds, but by enabling more accurate cross referencing, data matching, and pattern identification to a degree heretofore unobtainable within the everyday time, labour and economic limits in place. Taking advantage of experience with previous investigative stories, algorithms could be produced to trawl for combinations of data that have, in the past, been connected to newsworthy events. At the same time, the processing capabilities of computation devices can reveal relationships and connections that might otherwise be obscured by the volume of data at hand. The idea here is that an investment in codes and algorithms that can trawl data rapidly and effectively, particularly where these codes are scalable and open-source and can be easily tweaked for reuse, will greatly minimise the time and labour spent on tedious manual scanning and checking thus allowing more time for

verification, interviews and higher value activities. For example, the programming of ‘set-and-forget’ ‘watchdog’ algorithms that continuously monitor data-streams for emerging anomalies or interesting trends may enable news to break faster (Stenger 2008). By enabling journalists to make sense of complex data more quickly, computational journalism techniques produces datasets that can be easily verified and demonstrated, providing a robust and factual basis for investigative reporting that allows for the development of original stories more quickly and more thoroughly. For example, software can scan databases and social networks to identify and report patterns, which may be reviewed by journalists for story leads. Importantly, these leads would generally not have surfaced in any other way, and so increase opportunities for unique investigative journalism.

At one level, these debates sound similar to those surrounding *computer-assisted reporting (CAR)* in the late 1990s. Drawing upon the considerable enthusiasm among journalists for what the Internet could make possible, CAR promised a more scientific form of journalism, where journalists could cross-check information provided by key informants against publicly available data that was now easily accessible through the Internet, promising a new era of ‘precision journalism’ where truth-claims could be tested and backed up through thickets of verifiable data (Cox 2000). Why the current context of computational journalism differs from the scientific models promised by CAR in the late 1990s relates to two factors. First, CAR emerged in the era when the Internet was largely read-only: it offered masses of data, but the scope to comment on or interact with the data or with others using the data was severely limited. In the age of Web 2.0, the scope to communicate about and generate user communities around the data has been massively enhanced and expanded, so there is far more scope to test, contest and verify claims in the public domain. Second, CAR was a project to continue the high-modernist conception of journalism as a practice that could only be undertaken by those officially sanctioned as journalists, whether through employment at a news masthead, professional training in journalism schools, or informal inculcation into journalism culture.

Web 2.0 has dramatically expanded the scope for citizens of all kinds to engage in practices of public communication that are synonymous with journalism, whether it be through blogging citizen journalism, *WikiLeaks* or other means. Journalists now compete for attention in a much noisier digital public domain than was the case in the late 1990s when CAR was in its heyday, and it is a more open and interactive environment where, as Leonard Witt (2004) has observed, ‘public journalism’, where journalists and news editors could choose whether and how to engage with the public, has become ‘the public’s journalism’, where a new generation of media users are increasingly taking matters into their own hands. For journalists operating in this new environment, a certain level of digital literacy is needed in order to navigate multiple public and private databases and represent the complex information therein in interesting, relevant, and accessible ways, along with mastering the plethora of social media technologies needed to organise online communities around shared interests, issues and concerns (MEAA 2008, Doig 2008).

Contextual Factors Behind Computational Journalism

While the concept of computational journalism is not new, its potential value is increasing, and three major factors can be identified as driving it. First, there is the vastly increased amount of data that is publicly available, particularly information from government sources: regardless of whether that data is released through official or ‘underground’ channels such as *WikiLeaks*. Initiatives such as the U.S. Open Government Initiative, (www.data.gov), Britain’s data.gov.uk site, and the proposals of Australia’s Government 2.0 Task Force

(Department of Finance and Deregulation 2009) point to a growing trend towards making information held by government departments more openly accessible, and inviting ‘bottom-up’ re-use of this data under Creative Commons licences. Second, the combination of freely available Web 2.0 applications and the declining cost, greater ease of use and increasing power of data-mining software is promoting experimentation with publicly available data, “driven by a completely new upside down business model ... fuelled by freely available government and other data, [and] dependent on multiple actors some of whom may work for free, and which is often small scale and inexpensive” (Millard 2010). Finally, there is the more general explosion of myriad forms of online participation and engagement through a plethora of Web 2.0 and social media sites.

These are clearly not three separate sets of developments. Rather, they are three interconnected elements of a shift in the broader media ecology from top-down one-to-many mass communications media to more participatory and interactive social media cultures, that also impacts on politics and citizen engagement (Coleman 2005; Benkler 2006; McNair 2006; Jenkins 2006; Flew 2009). Computational journalism techniques may assist newspapers in particular in successfully adapting to this changing environment by generating new ways of balancing the need for quality, accuracy and originality in news in an environment where there is increasing speed of circulation of news, the need to lower costs in the face of reduced profit margins, and the desire to draw in online audiences as participants in the news process and not simply readers/consumers.

The race to the bottom of relying upon semi-automated news feeds, reporting fewer news stories and eliminating original in-depth investigative reporting altogether can be addressed, particularly for the established news brands with a reputation for quality and with loyal readerships (Flew *et. al.* 2009), in a relatively lightweight and low-cost manner. Mecklin (2009) argues that while investigative journalism that involves sorting through tens of thousands of documents looking for anomalies or incriminating evidence rarely produces news with enough appeal or utility value to make it economically viable in the disaggregated media world of the Internet, advances in computing can alter the economic equation by supplementing or even substituting for the slow and expensive human labour required to produce more useful outcomes. This is indicative of a shift in journalism practice from exclusive access to valued news sources to open access large-scale data sets as a means of gaining new information: a shift from a context of information scarcity to one of information abundance. And the idea of the evolving news story, which has a long history in journalism, is further enabled by an online environment that allows for individual stories to constitute what Bruns (2008) terms a *palimpsest* that changes over time in light of new information. For example, long-running, emerging, complex stories that utilise chronologies and timelines as reporting tools need no longer be re-presented in a linear chronological order, instead using intricate interactive interfaces that allow users to focus on time-tagged events, inter-related actions, and to turn on or off key actors or events to navigate and create personalised reconstructions of events. This could particularly appeal to the online-only news sites, which have small staff, low fixed costs, and a more rapid turnover of stories.

It needs to be borne in mind that established news mastheads now face new competitors in the online space. In addition to online-only news sites, there are content aggregators such as *Google News*, who particularly target the large number of *convenience users* of online news – that is, users who access online news by ‘push’ means (i.e. from other applications they were using), and are not strongly engaged with news as a media form (Flew, Spurgeon and Daniel 2009). Programs such as Google’s *Living Stories* (<http://livingstories.googlelabs.com/>) allow readers to use software tools to personalise their news, collating all versions of a news item into one article, and continually updating that article, while also providing interactive timelines on the news item. This suggests that the

alleged polarity between high-quality ‘masthead’ online news sites and low-quality aggregators may be a fiction that news organizations believe to their medium-term peril.

Benefits and Examples of Computational Journalism

Here, we show where computer journalism has enhanced reader experience and engagement with news while taking better advantage of the new information environment and reducing the cost of investigative reporting. Computational journalism can enhance user engagement and enable greater interaction with news media through new forms of communication and dissemination including online community and social networking tools, by making available to readers, interactive or participatory multimedia and data visualisation tools.

Computational journalism provides greater opportunities for collaboration and co-creation between professional journalists, citizen journalists, and their readers. Examples here include crowdsourcing and the co-reporting of news across platforms. Crowdsourcing, wherein groups of people work collaboratively via the internet on a single news item or part thereof, means that many people can spend a few minutes on low-level research that might take one person days to complete (Downie and Schudson 2009). *The Guardian* ‘MP Expenses’ investigation (<http://www.guardian.co.uk/politics/mps-expenses>) exemplifies the use of crowdsourcing techniques within investigative research projects. Triggered by the leak and subsequent publication by the Telegraph Group in 2009 of expense claims made by members of the United Kingdom Parliament over several years, *The Guardian* initiated in 2009 a thorough investigation into the expenses of Members of Parliament in the United Kingdom. They made publically available laboriously detailed data that readers could search by member, constituency or item, and send comments or questions to Guardian staff about the data. Guardian journalists investigated questionable claims, building stories around them. The flow of news items led to a government inquiry that found many members had incorrectly made claims. A key benefit to the Guardian was that no other media entity was able or prepared to replicate the study, enhancing the reputation of the Guardian for investigative journalism. This is despite the relatively low costs: Andersen (2009) writes that the software used for the investigation allegedly took a developer one week to build and an additional £50 to rent temporary servers. In this particular example of collaborative journalism, it was found that over 20,000 consumers participated (with a participation rate of 56 per cent), with 170,000 documents reviewed in the first 80 hours. The public’s contribution was recognised with the MP expenses page noting which users had reviewed most documents and found the most interesting items.

Another leading example in both crowdsourcing and co-reporting is that of WikiLeaks. Launched in 2006, WikiLeaks (www.wikileaks.org) is an international organization that publishes anonymously submitted document submissions and leaks otherwise unavailable to the public. The materials posted by WikiLeaks range from documents outlining Guantánamo Bay procedures, Sarah Palin's Yahoo email account contents, lists of forbidden or illegal web addresses for several countries including those to be banned under the Australian government’s proposed laws on internet censorship, e-mail correspondence between climate scientists leaked from the Climatic Research Unit of the University of East Anglia, video from an incident in which Iraqi civilians were alleged to have been killed by U.S. forces, and, perhaps most controversially, over 75,000 documents about the War in Afghanistan not previously available for public review (Associated Press 2010) and over 300,000 documents relating to the War in Iraq. The later leaks resulted in White House National Security issuing a statement saying the leaks were “irresponsible”, and that “the United States strongly

condemns the disclosure of classified information by individuals and organizations which could put the lives of Americans and our partners at risk, and threaten our national security" (USCC 2010).

In discussing the accuracy of the documents released, WikiLeaks states that it has never released a misattributed document and that documents are assessed before release. The site's FAQ states that: "the simplest and most effective countermeasure is a worldwide community of informed users and editors who can scrutinize and discuss leaked documents" (wikiLeaks.org 2007). WikiLeaks claims that by making publicly available leaked documents to the scrutiny of a worldwide community of informed 'wiki editors', they are able to offer a more exacting scrutiny than any media organisation could provide.

The co-reporting of news overcomes the economic and labour constraints faced by news providers who cannot source first-hand information from anywhere at any time. Journalists may use eyewitness reportage and recordings from localised citizens to provide more relevant news coverage. For example, people may upload photos of an event, sharing information, and confirm journalist enquiries. However, eyewitness citizens are often not journalists and are unequipped or unskilled for the activities needed to verify facts and to make sense of events in a journalistic manner. Skoler (2009) argues that collaboration between journalists and readers may help to build ongoing partnerships, lead to a greater understanding of what issues readers find relevant, and uncover genuine experts or people with first-hand experience that can inform the story. Twitter is one example of a computational tool that has been recently embraced by journalists as being capable of both promoting stories amongst peers and followers, and serving as a mechanism for interacting with followers to gauge opinion and verify information.

Myers (2009) argues that enhanced visualisation tools and interactive multimedia and graphics enable journalists to present key themes of investigations in more powerful and useful ways. Lists of data need no longer be presented as text-based or tabular information, but as dynamic, compelling and manipulable charts, maps and diagrams that allow readers to follow their interests and uncover meaning (Silfry 2009). For example, *LobbyLens* (GovHack 2009) correlates data sourced from twelve public agencies in a visual mashup showing federal government business links. The visual representation shows connections between government contracts, businesses, responsible ministers and other politician, and lobby groups. Clicking on a governmental department, minister, a business, and so on, shows links with other entities. A screenshot of the *LobbyLens* visualisation 'network graph' is provided in Figure 1 below.

(FIGURE 1 GOES ABOUT HERE)

While such mashups are dynamic and interactive they are often presented as unfiltered information, as opposed to a news item that identifies and expands upon links of interest and includes commentary from affected persons and credible experts. Nevertheless, data visualisations may provide a platform from which professional and citizen journalists can identify further news items. Indeed, even the most clear and useful data visualisations require some element of sense-making, from basic contextualisation to more in-depth explanation or narration. For example, post-poll election figures offer little sense without some explanation of the underlying reasons the results were returned.

Another example of where computer journalism draws on data visualisation comes with the proliferation of consumer-level Geographic Information System (GIS) and accompanying mapping programs, such as Google Maps, that enable the layering of variable data over maps to show clusters or patterns of localised information. While this means that location specific information need no longer be limited to summarised tables for print media,

it also means that the huge amount of detail and interactivity that are otherwise lost in static presentations remain catered for. One often cited example of such an application is the web service *EveryBlock* (<http://www.everyblock.com>). Covering 12 American cities with another 4 in beta, *EveryBlock* users can gather civic and other Web-based information such as police reports, traffic updates, building permits, news articles, blog posts, consumer reviews, photo and video connected to an immediate geographic location. Using this time-stamped and geo-tagged data, professional and amateur journalists can write articles personalised to user neighbourhood and interests, collating, for example, crime, transport and city health data with user recommendations and reviews to write articles on the best and safest eating and drinking within a specified distance. An early example of journalists using mapping can be seen in Dedman's (1988) *Colour of Money* articles (<http://powerreporting.com/color/>). Dedman's project involved overlaying income data and data derived from bank loans over maps of the city of Atlanta. The data showed that while Atlanta's banks and savings and loan institutions made loans in the poorer white neighbourhoods of Atlanta, they did not lend in middle-class or more affluent black neighbourhoods (Dedman 2008). Consequently, the U.S. Federal Home Mortgage Disclosure Act was expanded to provide more information to the public on the pattern of activity by all mortgage lenders. Contemporary visualisation tools, mapping software and databases would aid such investigative reporting immensely.

Developing Computational Journalism Capacity

There are a range of issues and challenges presented by the adoption of computational journalism in news organisations, notably significant software and technology start-up costs and the need for greater integrity measures in data-based reporting. Significant start-up costs may be reduced if the software can continue to be used. Alternatively, journalists and other news providers could share costs in partnership with third-party software developers. Recently, the Australian Government 2.0 Task Force has been instrumental in driving forward this agenda through its sponsorship of the *Mashup Australia* contest for innovative online applications that build on government data made available through the data.australia.gov.au portal (<http://mashupaustalia.org>), and its funding of a range of other research and development activities (<http://gov2.net.au/projects/>). Another example can be seen in the Sunlight Foundation (<http://www.sunlightfoundation.com>), a non-profit, nonpartisan organisation that has arisen in the light of the plethora of U.S. data made publicly available under initiatives of greater government openness and transparency. Aimed at revealing "the interplay of money, lobbying, influence and government in Washington in ways never before possible (Sifry 2009)," the Sunlight Foundation has been involved in both the creation of freely available tools and websites that enable individuals and communities to access and engage with government information, along with offering training to journalists and civic-minded citizens on how to use this data in local reporting.

Associated Press (AP) is using computational journalism to update and renew the historical aggregation role of the news agency. It is also taking an open source approach to leading the development of computational journalism capacity for news media, called *Overview* (Stray 2010; 2011). This initiative seeks to provide user-friendly interfaces for dealing with massive volumes of documents as well as the logistics of crowd sourcing and co-reporting. Other media participants in the project include *The Guardian*, and *The South China Morning Post*. The impetus for the project came from the WikiLeaks 2010 release of hundreds of thousands of documents relating to the wars in Iraq and Afghanistan. *Overview* relies on Document Cloud (<http://www.documentcloud.org/home>) to store and annotate documents. It then applies algorithms for keyword searches and frequency analyses to

automate the abstraction and visualisation of documents to into clusters of similar documents. The technique was successfully tested in a pilot analysis of WikiLeaks Iraq War documents for December 2006. It was used to organise documents into an open-ended number of clusters. By visualising patterns and relationships within and between clusters it was possible for journalists to quickly identify ‘newsworthy’ themes for closer investigation, including causes of death ranging from illegal killings to deaths and injuries associated with explosive hazards, This manipulation of data made the task of identifying useful starting points for deeper reading and closer investigation (by journalists, citizen journalists and readers alike) more manageable and purposeful.

Computational journalism often follows the procedure of finding suitable sets of data for analysis, making sense of that data, and representing this understanding in a way that is interesting, accessible and newsworthy. Often datasets are only imminently newsworthy when they are linked with other datasets (for example, mortality data with mapping data, public works data with parliamentary representative data, public health data with seasonal data and so on). Many of these valuable datasets are publically available through government and civil society open data programs such as data.gov or data.gov.uk or sites like WikiLeaks and OpenLeaks. Cohen *et. al* (2011) write that a problem often cited by journalists is that while interesting and newsworthy data might be readily available, it is often in formats such as .pdf files that make it both time consuming and labour intensive to glean data from. And while developments have been made in gathering (scraping) data from such sources (for example, the web-based Scraperwiki service <http://scraperwiki.com>), journalists often have to rely on crowd-sourcing to gather usable data.

However, while crowd-sourcing may be employed to assist journalists in facilitating fact-checking, it can also create problems for journalists as levels of expertise in analysis, levels of professionalism, user availability, commitment, flexibility, and reliability, and competing user interests can make it difficult to predict outcome quality and accuracy. Also, the logistics involved in the disaggregation of data-based tasks into component sets and the coordination of crowd-sourcing resources may place additional time and resource constraints on journalists. Nevertheless, quality control of data-based tasks might be afforded through ubiquitous human computing - the employment of professional third-party crowd-sourcing services such as CrowdFlower, Samasource, Feelancer.com, or the Amazon Mechanical Turk (Cohen *et. al* 2011). A ranking or rating system of these services would be a valuable tool for journalists working within this area.

Computational journalism can only adhere to traditional investigative journalistic principles if the datasets used can be shown to withstand rigorous and open fact-checking. And while crowd-sourcing proved to be an effective and valuable resource in this regard for *The Guardian's WikiLeaks* data projects, the value of these projects lay in how the Guardian visually represented that data. There is now a plethora of free online tools allowing non-expert users to consume or create meaningful images from large amounts of information for the purpose of insight including Google Labs' Public Data Explorer (<http://www.google.com/publicdata/home>) and IBM's Many Eyes (<http://www-958.ibm.com/software/data/cognos/manyeyes/>).

Data visualizations and graphics can help both readers and journalists cut through dense information in an efficient way. Cohen (in Myers 2009) suggests that two of the most important elements in reporting a complex story are place and time, elements that are often difficult to glean from large amounts of raw information, particularly when in combination. Data visualisations can not only show journalists and readers where to find typical subjects or when a particular event occurred, but can enable journalists to highlight frequency, groupings by age, by event, by location, or some other variable element or combination of elements. Such visualisations may be used only for a journalist's self understanding, or to help focus or

refine a story, while others will be more suitable for publication. For example, complex stories that develop over a period of time may benefit from an interactive chronology that enables journalists to zoom in and out on specific periods of time or to turn on and off various types of entries or variables. The same software could help readers navigate the story, allowing them to back-track if they arrived late, or understand the movement and action of key elements over a period of time beyond a particular article.

Yet while data visualisation is a practical and constructive tool within the grasp of the pro-am journalist and blogger, a more useful set of skills, however, is the professional journalist's ability to recognise, frame and contextualise the most interesting and important news stories hidden within the data. Often the most *obvious* instantiations of data are by no means the most *interesting*, and data represented in a particular way may not only be represented differently, but may generate new stories when the parameters and schema are adjusted. Indeed, as newsrooms continue to seek greater online audience participation, they may find that readers themselves are best positioned to play with datasets and create their own understanding from the data being presented while the journalist simply verifies the datasets and provides a broader contextualising framework. Coupled with an estimable data visualisation tool, such practice would allow readers to humanise or localise what may otherwise be large, incomprehensible sets of data.

Conclusion

The strength of data derived journalism relies on the integrity, quality and reliability of the data available: a central concern for both journalism and information technology. Statistical anomalies, a lack of sense, misinterpretations, conflicting data standards, incomplete data, skewed results or malignly altered data pose serious problems. The misuse of data includes exposing or trading private data from data sets that should be aggregated and de-identified. Mecklin (2009) offers an example in which the United States government mined multiple databases to identify signals of terror activities, and highlights the breaches of privacy this incurred. The risk of data misuse necessitates vigilant oversight for data gathering practices. Perhaps with a view to this, the United Kingdom government has moved to increase penalties for data misuse (Ministry of Justice 2009). Nevertheless, while any additional need to verify, qualify and explain data projects will add to the time and labour costs associated with developing a story, computational programs that verify data coupled with crowdsourcing could assuage these costs.

Ultimately the utility value of computational journalism comes when it frees journalists from the low level work of discovering and obtaining facts, thereby enabling greater focus on the verification, explanation and communication of news. Such an understanding serves to dissolve the illusion that news providers employing computational journalism can automatically deliver *better* news to their readers simply because they are able to move more information about at faster speeds, and from more remote locations. In other words, computational journalism has less to do with systems that transmit data and information only as a commodity. Computational journalism, like journalism *per se* is a constructive, meaning-making enterprise. In Dervin's (2003) sense-making methodology, theories of communication based on traditional transmission models are shown to be unable to account for the constructive and subjective enterprise that communication really is. Dervin argues that sense-making looks at the *hows* of communicating; that is, how individuals and collectives define situations, how they bring past experiences to bear, how they make connections, and so forth. Sense-making reconceptualises the making of facts assumed to be real as only one of many ways we make sense of our world. Alongside the making of facts consensus making, negotiation, power-brokering, defining, uncovering, emoting, cognising, and so forth are a

few of the many ways that we make, reinforce, challenge, resist, alter, and reinvent meaning (141-142).

Regardless of the changes within the broader media ecology, the demand for journalists to inform the public and hold Governments accountable remains strong and investigative journalism is seen as critical to the maintenance of democracy (Knight Commission 2009; Bunz 2009). For journalists involved in the making of meaning, computational tools serve to extend and supplement rather than supplant their skills. As Mecklin (2009, 4) writes, “there is no person more important to the general civic health — and yet more consistently under-rewarded in financial and social status terms — than the quality investigative reporter at a local news organization”. Computational journalism may assist news providers by generating new ways of producing quality news at greater speed and with reduced costs to active and participative audiences.

While this paper has provided some oversight into the steps and processes journalists might undertake in developing stories it is not meant to present as a thorough “how-to” guide. Rather, computational journalism brings a major new development to the field of journalism, and it is not, by any means, “business as usual”. Computational journalism demands not only a certain level of new ICT skills, capacities and literacies of journalists, but a new understanding of *how* journalists can work with, and in, the new economies of distributed and co-creative production. Computational journalism heralds an expanded vision wherein the coordination role of the journalist expands to manage the myriad of distributed micro-tasks surrounding data-based reportage; where the co-creation of news stories incorporates both information and resources crowdsourced from a distributed value net rather than flowing from a discrete value chain; and where unknown data outcomes mean that journalists may not know the shape of an unfolding news until the data itself is thoroughly interpreted, analysed, sorted, and processed to produce information about discernable trends, patterns and summaries that are not only accurate and meaningful, but add real value to journalistic knowledge production.

References

- Aarts, Emile, Ter Horst, Herman, Korst, Jan and Verhaegh, Wim (2006) “Computational Intelligence”, in: Emile Aarts and Jose L. Encarnação, (Eds), *True Visions: The emergence of Ambient Intelligence*, New York: Springer, pp. 245-273
- Andersen, Michael (2009) “Four crowdsourcing lessons from the Guardian’s (spectacular) expenses-scandal experiment”, *Nieman Journalism Lab*, 23 June, <<http://www.niemanlab.org/2009/06/four-crowdsourcing-lessons-from-the-guardians-spectacular-expenses-scandal-experiment/>>, accessed 10 May 2010.
- Associated Press (2010) “WikiLeaks to publish new documents”, *The Age*, 8 August, <<http://news.theage.com.au/breaking-news-world/wikileaks-to-publish-new-documents-20100808-11prq.html>>, accessed 10 August 2010.
- Benkler, Yochai (2006) *The Wealth of Networks: How Social Production Transforms Markets and Freedom*, New Haven: Yale University Press.
- Boden, Margaret (1985) “The social impact of thinking machines”, in books: in: Tom Forester (Ed), *The information technology revolution*, Oxford: Basil Blackwell, pp. 95-103.
- Brown, John Seely and Dugid, Paul (2000) *The social life of information*, Boston: Harvard Business School Press.
- Bruns, Axel (2008) *Blogs, Wikipedia, Second Life and Beyond: From Production to Producership*, New York: Peter Lang.
- Bunz, Mercedes (2009) “Rupert Murdoch: ‘There’s No Such Thing As A Free Story’”, *The Guardian*, 1 December, <<http://www.guardian.co.uk/media/2009/dec/01/rupert->

- [murdoch-no-free-news](#)>, accessed 8 December 2010. Cohen, Sarah, Li, Chengkai, Yang, Jun and Yu, Cong (2011) "Computational Journalism: A Call to Arms to Database Researchers". In *Proceedings of CIDR 2011*, pp.148-151 .
- Coleman, Stephen (2005) "New Mediation and Direct Representation: Reconceptualising Representation in the Digital Age", *New Media and Society* 7 (2), pp. 177-198.
- Cox, Meli (2000) "The Development of Computer-Assisted Reporting", paper presented to the Association for Education in Journalism and Mass Communication, Chapel Hill, NJ: University of North Carolina.
- Dedman, Bill (1988) "The Color of Money", *Power Reporting*, 1-4 May, <<http://powerreporting.com/color/>>, accessed 2 December 2010..
- Department of Finance and Deregulation (2009) , <<http://www.finance.gov.au/publications/gov20taskforcereport/index.html>>, accessed 12 August 2011.
- Dervin, Brenda, Foreman-Wernet, Lois with Lauterbach, Eric (2003) *Sense-making methodology reader: Selected writing of Brenda Dervin*, New Jersey: Hampton Press.
- Doig, Stephen K (2008) "Reporting With the Tools of Social Science", *Nieman Reports*, Spring, <<http://www.nieman.harvard.edu/reportsitem.aspx?id=100075>>, accessed 2 January, 2010..
- Downie, Leonard and Schudson, Michael (2009) "The Reconstruction of American Journalism", *Columbia Journalism Review*, 19 October, <http://www.cjr.org/reconstruction/the_reconstruction_of_american.php>, accessed 14 December.
- Flew, Terry (2009) "Democracy, Participation and Convergent Media: Case Studies in Contemporary Online News Journalism in Australia", *Communications, Politics and Culture*, 42 (2), pp. 87-109.
- Flew, Terry, Spurgeon, Christina and Daniel, Anna (2009) *Audience and Market Foresight: Consumer Use of Digital News and Information in Australia*, Sydney: Smart Services Co-operative Research Centre, <http://www.smartservicescrc.com.au/PDF/Audience_Market_Foresight_Vol1_Pub_Feb2009.pdf>, accessed 10 December 2010.
- Govhack (2009) LobbyLens, <<http://team7.govhack.net.tmp.anchor.net.au/networkgraph.php>>, accessed 12 December 2010.
- Hamilton, James T. and Turner, Fred (2009) "Accountability through Algorithm: Developing the field of computational journalism", Report from Center For Advanced Study in the Behavioural Sciences, Summer Workshop 27-31 July, <http://dewitt.sanford.duke.edu/images/uploads/About_3_Research_B_cj_1_finalreport.pdf>, accessed 11 December 2010.
- Jenkins, Henry (2006) *Convergence Culture: When Old and New Media Collide*, New York: New York University Press.
- Klein, Gary, Moon, Brian and Hoffman, Robert R. (2006) "Making Sense of Sense-Making", *IEEE Intelligent Systems* 21 (4), pp. 70-73.
- Knight Commission (2009) *Informing Communities: Sustaining Democracy in the Digital Age*, <<https://secure.nmmstream.net/anon.newmediamill/aspen/kcfinalenglishbookweb.pdf>> accessed 30 October 2010..
- McNair, Brian (2006) *Cultural Chaos: Journalism, News and Power in a Globalised World*, London: Routledge.
- Mecklin, John (2009) "Deep Throat Meets Data Mining", *Miller-McCune*, 10 January, <<http://www.miller-mccune.com/media/deep-throat-meets-data-mining-875>>, accessed

- 20 August 2010.
- Media Entertainment and Arts Alliance (2008) *Life in the Clickstream. The Future of Journalism*, <http://www.alliance.org.au/documents/foj_report_final.pdf>, accessed 20 January 2010,
- Meyer, Philip (2002) *Precision Journalism: A Reporter's Introduction to Social Science Methods*, (4th Edition) Lanham, MD: Rowman & Littlefield.
- Millard, Jeremy (2010) "Government 1.5 – is the bottle half full or half empty?", *European Journal of e-Practice* 9, pp. 1-16.
- Miller, John H. and Page, Scott E. (2007) *Complex Adaptive Systems: An Introduction to Computational Models of Social Life*, Princeton, NJ: Princeton University Press.
- Minsky, Marvin (1968) *Semantic information processing*, Cambridge: MIT Press.
- Ministry of Justice (2009) "The Knowing or Reckless Misuse of Personal Data Introducing Custodial Sentences", Consultation Paper CP22/09, <<http://www.justice.gov.uk/consultations/docs/data-misuse-increased-penalties.pdf>>, accessed 11 January 2010.
- Myers, Steve (2009) "Using Data Visualization as a Reporting Tool Can Reveal Story's Shape", *Poynter*, 14 April <<http://www.poynter.org/column.asp?id=101&aid=161675>>, accessed 10 July 2010.
- Sifry, Micah (2009) "A See-Through Society. How the Web is opening up our democracy", *Columbia Journalism Review*, 15 January, <<http://sunlightfoundation.com/press/articles/2009/01/15/see-through-society/>>, accessed 12 July 2010..
- Skoler, Michael (2009) "Why the News Media Became Irrelevant-And How Social Media Can Help", *Nieman Reports* 63(3), <<http://www.nieman.harvard.edu/reportsitem.aspx?id=101897>>, accessed 14 October 2010.
- Stenger, Brad (2008) "Wiring Journalism 2.0", *Columbia Journalism Review*, February, <http://www.cjr.org/the_observatory/wiring_journalism_20_1.php>, accessed 20 October 20 2009.
- Stray, Jonathon (2010) "A full-text visualization of the Iraq War Logs", *Jonathon Stray* (blog), <<http://jonathanstray.com/a-full-text-visualization-of-the-iraq-war-logs>>, accessed 28 July 2011.
- Stray, Jonathon (2011) "Investigation thousands (or millions) of documents by visualizing clusters", *Overview*, <<http://overview.ap.org/>>, accessed 28 July 2011.
- Taylor, Megan (2010) "How Journalists Can Incorporate Computational Thinking into Their Work", *Poynter*, 3 August, <<http://www.poynter.org/column.asp?id=31&aid=187439>>, accessed 10 September 2010.
- United States Central Command (2010) "Statement of National Security Advisor Gen. James Jones on WikiLeaks", 25 July, <<http://www.whitehouse.gov/the-press-office/statement-national-security-advisor-general-james-jones-wikileaks>>, accessed 10 August 2010.
- WikiLeaks (2007) "FAQ-EN", <<http://wikileaks.org/faq-en>>, accessed 10 August 2010.
- Wing, Jeannette M. (2006) "Computational Thinking", *Communications of the ACM* 49(3), pp. 33-35.
- Witt, Leonard (2004) "Is Public Journalism Morphing into the Public's Journalism?", *National Civic Review* Fall, pp. 49-57.

Authors contact details:

Terry Flew

Email: t.flew@qut.edu.au

Address: Creative Industries Faculty, Queensland University of Technology, Musk Avenue,
Kelvin Grove, 4059, Brisbane Australia

Christina Spurgeon

Telephone (for proofing queries only – not for publication): + 61 405 132 013

Email: c.spurgeon@qut.edu.au

Address: Creative Industries Faculty, Queensland University of Technology, Musk Avenue,
Kelvin Grove, 4059, Brisbane Australia

Anna Daniel

Email: annamdaniel@gmail.com

Adam Swift

Email: a.swift@qut.edu.au

Address: Creative Industries Faculty, Queensland University of Technology, Musk Avenue,
Kelvin Grove, 4059, Brisbane Australia

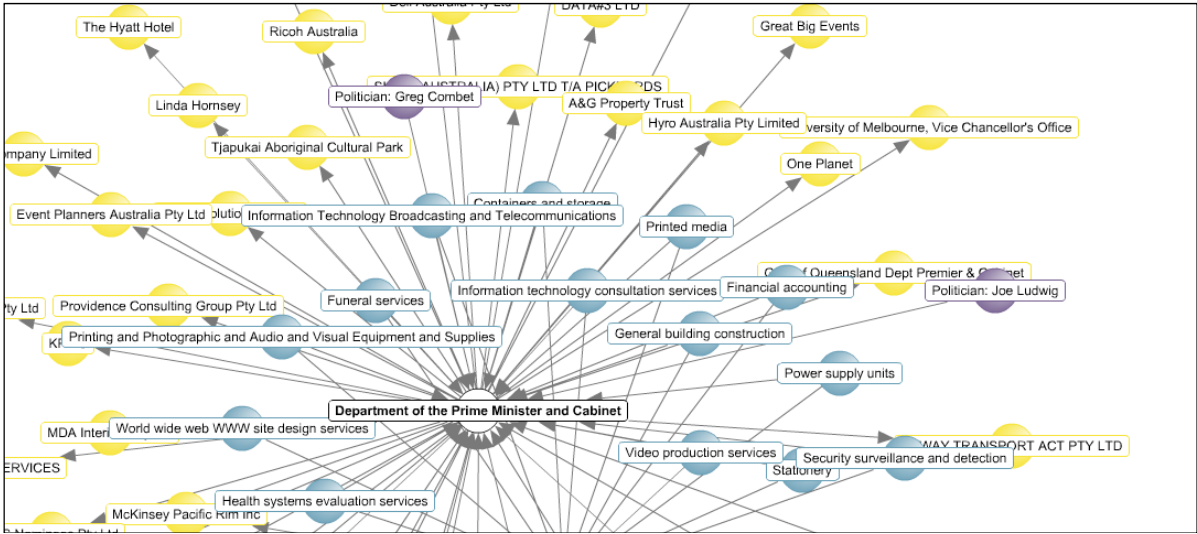


Figure 1: LobbyLens

Authors Bioblurb

Terry Flew

Email: t.flew@qut.edu.au

Address: Creative Industries Faculty, Queensland University of Technology, Musk Avenue, Kelvin Grove, 4059, Brisbane Australia

Terry Flew is Professor of Media and Communications in the Creative Industries Faculty at the Queensland University of Technology, Brisbane, Australia. He heads the New Media Services work programs of the Smart Services Co-Operative Research Centre, and is a Chief Investigator in the ARC Centre of Excellence for Creative Industries and Innovation. He is currently seconded to the Australian Law Reform Commission to lead a review of the national classification system. He is the author of Australia's leading new media textbook, *New Media: An Introduction* (Oxford, 2008 - third edition), and *Understanding Global Media* (Palgrave, 2007).

Christina Spurgeon

Email: c.spurgeon@qut.edu.au

Address: Creative Industries Faculty, Queensland University of Technology, Musk Avenue, Kelvin Grove, 4059, Brisbane Australia

Dr Christina Spurgeon is a Senior Lecturer in Journalism, Media and Communication in the Creative Industries Faculty at the Queensland University of Technology, Brisbane, Australia. Co-creative media production practices, and the implications for media and communication industries and institutions, are her present research focus. Dr Spurgeon's book, *Advertising and New Media*, was published by Routledge in 2008.

Anna Daniel

Email: annamdaniel@gmail.com

Anna is a commercial researcher and business manager and her projects have given her particular insight into the digital media and entertainment sectors. Anna is involved in a national research project that explores creative industries in suburban locations. She has worked as a research associate in the Creative Industries Faculty at the Queensland University of Technology and has previously worked in corporate and government strategy and research management positions, including at PricewaterhouseCoopers, Accenture, Commonwealth Funds Management, Federal Government Departments and public radio sector.

Adam Swift

Email: a.swift@qut.edu.au

Address: Creative Industries Faculty, Queensland University of Technology, Musk Avenue, Kelvin Grove, 4059, Brisbane Australia

Dr Adam Swift is a research associate in the Creative Industries Faculty at Queensland University of Technology in Brisbane, Australia, and is currently part of a Smart Services Cooperative Research Centre team exploring social trends in online news and information services. Swift has published on a range of topics associated with user engagement with New Media Technologies.