



**Queensland University of Technology**  
Brisbane Australia

This is the author's version of a work that was submitted/accepted for publication in the following source:

Flew, Terry, Daniel, Anna, & Spurgeon, Christina L. (2010) The promise of computational journalism. In McCallum, K (Ed.) *Media, Democracy and Change: Refereed Proceedings of the Australian and New Zealand Communications Association Annual Conference*, Australia and New Zealand Communication Association, Canberra, ACT, pp. 1-19.

This file was downloaded from: <http://eprints.qut.edu.au/39649/>

**© Copyright 2010 please consult the authors**

**Notice:** *Changes introduced as a result of publishing processes such as copy-editing and formatting may not be reflected in this document. For a definitive version of this work, please refer to the published source:*

# The promise of computational journalism

**Anna Daniel, Terry Flew & Christina Spurgeon**

Anna Daniel is a commercial researcher and business manager and her projects have given her particular insight into the digital media and entertainment sectors. As a research fellow at the School of Journalism and Australian Studies at Monash University, Anna is involved in a national research project that explores creative industries in suburban locations. As a research associate in the Creative Industries Faculty, Queensland University of Technology, Anna is part of a team exploring trends in digital media. She 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. Anna has submitted for a PhD in Business, with a thesis that explored emerging business models in the music sector.

Terry Flew is Professor of Media and Communications in the Creative Industries Faculty at the Queensland University of Technology, Brisbane, Australia. He is the author of Australia's leading new media textbook, *New media: An introduction* (Oxford, 2008, 3rd edn.), and *Understanding global media* (Palgrave, 2007). From 2006 to 2009, he headed an ARC Linkage project on citizen journalism in Australia, with the Special Broadcasting Service, Cisco Systems Australia and The National Forum as industry partners. He is also leader of an ARC Discovery project on creative suburbia with a research team from Queensland University of Technology and Monash University. 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.

Christina Spurgeon lectures in Journalism, Media and Communication in the Creative Industries Faculty at Queensland University of Technology. She is presently assisting with a faculty-based capacity-building project in the development of co-creative media research and continuing professional education services. She is also a Chief Investigator on an ARC Discovery project, *New media voices in the Australian values debate*, and author of *Advertising and new media* (Routledge 2008).

## **Abstract**

*There are at least four key challenges in the online news environment that computational journalism may address. Firstly, news providers operate in a rapidly evolving environment and larger businesses are typically slower to adapt to market innovations. News consumption patterns have changed and news providers need to find new ways to capture and retain digital users. Meanwhile, declining financial performance has led to cost cuts in mass market newspapers. Finally investigative reporting is typically slow, high cost and may be tedious, and yet is valuable to the reputation of a news provider.*

*Computational journalism involves the application of software and technologies to the activities of journalism, and it draws from the fields of computer science, social science and communications. New technologies may enhance the traditional aims of journalism, or may require "a new breed of people who are midway between technologists and journalists" (Essa, cited in Mecklin, 2009, p. 3). Historically referred to as "computer assisted reporting", the use of software in online reportage is increasingly valuable due to three factors: larger datasets are becoming publicly available; software is becoming sophisticated and ubiquitous; and the developing Australian digital economy.*

---

*This paper introduces key elements of computational journalism—it describes why it is needed; what it involves; benefits and challenges; and provides a case study and examples. Computational techniques can quickly provide a solid factual basis for original investigative journalism and may increase interaction with readers, when correctly used. It is a major opportunity to enhance the delivery of original investigative journalism, which ultimately may attract and retain readers online.*

---

## **Introduction**

Computational journalism can be broadly defined as the application of computer science techniques to the activities of journalism. The potential for computational journalism is driven by:

1. the increasing transparency of government information;
2. the ubiquity and power of software; and
3. the developing Australian digital economy.

Its use may:

1. increase the depth of quality original investigative reporting;
2. differentiate a digital masthead from competitors;
3. accelerate the process of journalism from news source to delivery;
4. be more easily implemented by mass market news providers; and
5. provide a factual basis for analysis, which may minimise the risk of incorrect reporting.

Ultimately these benefits may attract and retain online readers.

This paper is written from a media and social research perspective as distinct from a technology perspective. Relevant terms to computational journalism are provided in the glossary and examples of computational journalism in practice have been provided as an appendix.

## **The need for computational journalism**

There are at least four key challenges in the online news environment that computational journalism may address, and these will be discussed. Firstly, news consumption has changed in the digital realm, consumers are no longer a “mass”; instead they use news in different ways and their niche interests are not being met by current mass market offerings. A consequence may be the journalism profession’s “increasing isolation from the needs of its readers and viewers”, according to Michael Skoler, executive director of the Center for Innovation in Journalism (Georgia Tech, 2008, p. 4). Skoler does not imply journalists become “social chats”, rather that they may identify and connect more deeply with those who have first hand experiences that may inform their research.

Daniel, Flew and Spurgeon (2009) identified three typologies of online news users in Australia:

1. access news conveniently or by default when online doing other things, e.g. news may “pop up” while they are on Facebook, or they are redirected to news sites when they log out of email sites;
2. are loyal to a masthead brand—they seek out the internet equivalent of their physical newspaper; or
3. actively customise their news online, e.g. use RSS feeds to filter news, and actively interact with news by commenting on or reusing news via blogs.

Convenience users comprised 60 per cent of their sample population; however, loyal and customising users may be more valuable to an online news site. These users appreciate quality news content and are more likely to interact heavily with news media. How may newspaper businesses capture and retain these audiences?

Declining financial margins at traditional news providers have resulted in cost cuts. Investigative reporting within many news organisations has become a luxury (Ide & Vashisht, 2006) and been hardest hit by cost cuts (Steiger, 2009). Fewer journalists are reporting less news in fewer pages (Downie & Shudson, 2009) and news organisations rely increasingly upon newsfeed content. Perhaps journalists feel some pressure to “do more with less” (Media Entertainment and Arts Alliance, 2008). However, consumers expect original news and deep analysis from quality mastheads. Similarly, local news is highly valued, and this tends to be original rather than received from syndicated newsfeeds. How can journalists maintain quality reporting standards?

Thirdly, investigative journalism is typically slow to research, high cost, and some elements of the research activity may be tedious (e.g. reviewing documents for evidence) (Steiger, 2009; Mecklin, 2009). Researchers trawl through datasets of different media—e.g. excel files, PDFs, HTML, wikis, word documents, video and multimedia, and social networks. Yet investigative journalism is critical to the maintenance of a credible news brand, and it differentiates them from tabloid sites that may rely upon content from low cost commodity newsfeeds. Tabloid sites appear to be damaging consumer perceptions of quality in the digital news sector (Daniel, Flew & Spurgeon, 2009). How can online news suppliers streamline costs and increase the efficiency and speed of original investigative journalism?

Fourthly, the velocity of change in the technology environment has threatened the newspaper business model. City University Professor of Journalism Rou Greenslade has observed that investment gurus such as Warren Buffett argue the newspaper sector was complacent about the internet (Greenslade, 2009). On the real estate of the internet, users make less distinction between established mastheads and niche providers, or between newspapers and other media. Legacy news providers now compete online with niche “online only” news sites, which may be smaller, nimble and have less fixed costs. David Penberthy at the Punch reflects on this difference:

In the past 12 months at my job I have gone from working full-time at a large newspaper with a staff of about 200 to working full-time at a website with a staff of four. The media is nowhere near as prone to the meeting culture as other businesses but, even so, on the newspaper we would have meetings every day, sometimes four or five of them, at the website we would have one a month.

And it has been illuminating to see how much more you can get done having an irregular but necessary meeting with a small number of like-minded people, rather than

a series of daily meetings with a large number of people, many of whom would rather be somewhere else (Penberthy, 2010).

The capacity of mass market news providers to compete may be hindered by their size, which makes them slow to adapt to market changes. They traditionally appeal to a broader mass audience (albeit typically with certain ideological slants in readership), which makes it more difficult to respond to market dynamics quickly. However their size and mass market appeal may also be an advantage. In the short term, smaller news entities may not have the capacity to employ or train staff who specialise in programming or software development, whereas larger media companies may already have this expertise in-house. Some major newspapers in the United States are actively recruiting software developers (for example, in June 2009, the *New York Times* advertised on monster.com for a range of journalism-inclined software developers). Secondly, mainstream news providers reach a mass audience and this can be an advantage in “crowd sourcing” background research (to be discussed later as a computational journalism technique). These are both short-term advantages as: software becomes more powerful, lower cost and easier to use by generalists; journalists increasingly learn to program (Villano, 2009); and startups increase their audience numbers.

Mass market news providers also sense some threat from larger technology competitors—e.g. content aggregators. *Google news* presents “news” items with minimal intervention from journalists, using automated clustering and indexing of news feeds. It acts as a directory to news content, as opposed to a content creator. At the other end of the scale, consumers may personalise their news and tailor the delivery of news to their interests.

Computational journalism offers an opportunity to improve the quality and efficiency of news reporting. Higher quality news reporting will retain “loyal” and “customising” consumers.

## **Computational journalism defined**

Computational journalism involves the application of software and technologies to the activities of journalism. It draws from the fields of computer science, social science and communications. Computational journalism is defined more formally by Hamilton and Turner (2009, p. 2) 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.

Irfan Essa says both journalism and information technology are concerned with information quality and reliability, and computational journalism brings technologists and journalists together so they can create new computing tools that further the traditional aims of journalism, or “we are talking about a new breed of people who are midway between technologists and journalists” (cited in Mecklin, 2009, p. 3).

Newsgathering corresponds closely to an area of computer science known as sense making—how to go from not understanding a problem to understanding it.

Computational techniques may enable journalists to make sense, to undertake analysis

and create original stories, faster 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.

The concept of computational journalism is not new, and older journalists may recall computer-assisted reporting techniques in the 1990s. But its potential value is increasing, driven by:

- (a) government policies to make available larger scale databases of data for scrutiny, to increase public service transparency (Singel, 2009);
- (b) the declining cost, increasing power and ease of use of data-mining and filtering software, and Web 2.0; and
- (c) the explosion of online public engagement and opinion—the proliferation and scale of user-generated content such as blogs, twitter feeds and social network content that may or may not contain information of interest (Sifry, 2009).

Ultimately as the ease of use and sophistication of software improves, consumers may directly use such tools to personalise their news—e.g. via Google’s Living Stories. Living Stories collates all versions of a news item into one article, and updates that article. Secondly it links to related news, opinion pieces, in-depth reports or organisations. Thirdly, Living Stories provides an interactive timeline on the news item (Google, 2010). The practice of computational journalism is best explained via examples, and a selection is supplied throughout this paper and as an appendix.

## **Benefits of computational journalism**

I think that this is much more a tool to inform reporters, so they can do their jobs better. (Sunlight Foundation’s Bill Allison cited in Mecklin, 2009, p. 3)

Benefits from greater application of computational techniques for news organisations include:

1. Improving the quality of news journalism by increasing the speed of sense-making from diverse sources. It identifies data patterns that may lead to original news.
2. Making more effective use of publicly available data. Consumers or niche news sites may not be in a position to fully exploit such databases, because they do not have the scale of resources available to do so, nor possess the expertise that investigative journalists can bring to the analysis of such data. For example, *The Guardian* initiated an investigation into the expenses of Members of Parliament via a crowd sourcing project, and no other media agency was able or prepared to replicate their study in the United Kingdom (described in Appendix 1).
3. Adding value to digital news by providing readers with tools by which they can check facts. Reporting using multimedia and interactive elements “will succeed by further engaging readers who are faced with an ever growing number of options for news” (Georgia Tech, 2008, p. 3).
4. Potentially minimising the time spent on tedious background research and fact checking. The time investment changes from one of manual scanning to that of creating codes and algorithms to structure, mine and report data, which can then be reused on other projects.

5. Increasing the speed of investigative research. This may allow more time for verification, interviews and higher value activities and ultimately it may enable news to break faster.
6. It may change the nature of news by new forms of communication and dissemination including social networking tools, interactive or participatory multimedia and data visualisation.
7. It may minimise the risk of potential legal actions from incorrect reporting, because it involves robust techniques from computer science and datasets to produce data that can be explained and demonstrated. It provides a factual basis for further investigative reporting.

Ultimately, computational techniques improve the depth and context of investigative reporting. Similarly, the use of visualisation tools may present news more powerfully, and journalists may be best able to communicate the key themes of investigations this way. For example, LobbyLens (GovHack, 2009) is a data mashup about federal government business links and it correlates data sourced from twelve public agencies. It then presents it to show connections between government contracts, business details, politician responsibilities, lobbyists, clients of lobbyists and the location of these entities. Users can click on a department, person, business, or contract and it will show links with other entities. A screenshot of the visualisation “network graph” is provided in figure one below.

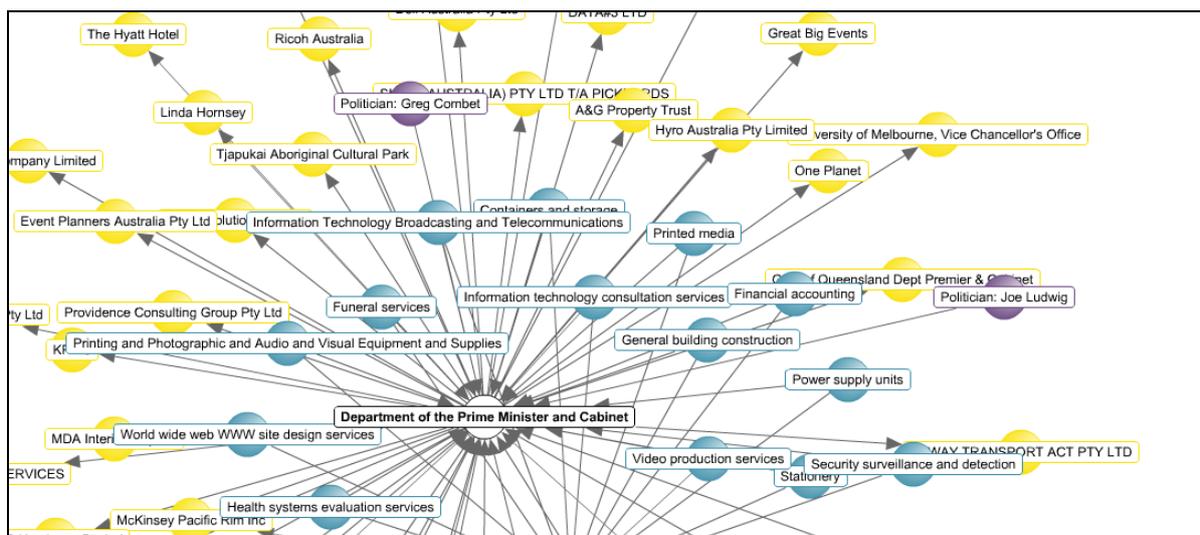


Figure 1 LobbyLens

The mashup is dynamic and interactive, but is presented as information, as opposed to a news item that may identify and expand upon links of interest, include commentary from affected persons and credible experts. It may provide a platform from which journalists can identify news items, investigate further and present stories, or alternately consumers can use the mashup to identify news items, and then ask journalists to investigate.

## Techniques

Computational journalism borrows from computer science and social science methodologies. The following techniques are described in laypersons terms, and may be elucidated further by a computer scientist.

## **Statistical analysis**

Techniques include hypothesis testing, sampling and probability estimates. Some numbers follow Benford's Law, which means that in the distribution of first digits, ones outnumber twos, which outnumber threes, and so on. Work in forensic accounting shows that when people "fudge their numbers" they forget to do it in a way that replicates Benford's Law. This means that analysis of first digits is a way to check the accuracy of self-reported data. Similarly, if there is no change in the pattern of data reports over time it may suggest the data is not representative or underreported (Hamilton, 2009).

## **Regression analysis**

Many newspapers have used regression analysis to examine the merits of standardised testing as a means of assessing the quality of education in public schools. The *Dallas Morning News* broke a series of stories about schools that were cheating on standardised tests in order to improve their overall ranking. The story began with a regression analysis that showed these schools (quite literally a needle in a haystack of data) to be outliers (Doig, 2008).

## **Correlation and matching**

Gathering, cross checking and looking for matches in datasets to identify patterns—for example, LobbyLens (described above) depicts the relationships between political parties and lobbyists.

## **Visualisation, mashups and GIS**

Visualisation tools help journalists identify patterns quickly and tell their stories more clearly and powerfully, as seen in the LobbyLens example. The layering of data with maps allows clusters or patterns to be identified visually, as can be seen in the "Hurricane damage" example in Appendix 1. Visualisation software, mashups and other interactive multimedia and graphics may represent news more appropriately (Myers, 2009). Such tools may also facilitate interactions between readers and consequently contribute to the research project. Digital interactive storytelling may be facilitated in this way—for example, the Mediastorm (2010) project provides a selection of multimedia stories (as described in the Appendix).

## **Parsing**

Overlaying demographic data over other data such as bank loans—e.g. the 1998 "*Color of Money*" investigation (detailed in Appendix 1)—is now much easier and faster to undertake (Dedman, 2008).

## **Personalisation**

At the hyperlocal level, in locations that are too small or remote to sustain full-time journalists, algorithms could write stories based upon hyperlocal news. This may be perceived as low quality reporting, but could be supplemented by input from, and interaction with, community members. Recommender systems (e.g. ratings and comments) may also enhance the personalisation of local news.

## Co-creation

Between journalists and readers is another technique that is emerging in newsrooms. Co-creation may generate efficiency gains in the reporting of news, for example, hundreds of people can spend a few minutes on low-level research that might take one person days to complete. It includes:

1. crowd sourcing;
2. co-reporting of news between journalists and citizens across platforms; and
3. facilitating multimedia forms of citizen reporting, and citizen journalism.

Crowd sourcing techniques may be used within computational journalism projects. *The Guardian* “MP Expenses” investigation (referred to previously and described in Appendix 1) exemplifies the use of crowd sourcing techniques within investigative research projects. These techniques may help journalists to “build a partnership with the public . . . [to] understand what issues are relevant to people, and find the genuine experts with first-hand experience to inform the story” (Skoler cited in Georgia Tech, 2008, p. 4). The use of consumers by newsroom staff to assist with investigative research and reporting may increase as consumers become familiar with tools and are educated in their use. Journalists cannot be everywhere, but may use eyewitness reportage from consumers to enable more relevant news coverage (Georgia Tech, 2008).

## Challenges of computational journalism

The issues and challenges presented by the adoption of computational journalism in news organisations include:

1. Convincing decision makers of the potential of computational journalism. In theory it is quite dull, but in practice potentially provides innovative, original, quality investigative research. Some may deem it to be “far fetched” (Mecklin, 2009) or may dismiss it as a digital rendition of print. In a time of cost-cutting, businesses may be unwilling to divert resources into a new initiative (Villano, 2009). However computational journalism “draws a demographically different audience than print” (Georgia Tech, 2008, p. 2), possibly early adopters. These may be active “customisers” and “producers” (Bruns, 2008) with high expectations from online media and who may be more prepared to pay for services that address their expectations. The experiences of customisers may influence future growth rates of new media initiatives.
2. Cost and technical issues associated with starting up the computational journalism process may hinder its take-up. There may be significant up-front software costs, but once enabled, software can be reused in other ways quickly, so the process becomes easier. Software can be purchased or may be open source (students have prepared several in response to a competition, which can be viewed at Mashup Australia, as noted in Appendix 2).
3. Dirty data. This includes statistical anomalies such as double falsehoods, a lack of sense, misinterpretations or gaming of results, conflicting data standards or incomplete data. Some computational journalism projects may involve crowd-sourcing which may be gamed by vested interests. Also software may be used that

increases hits of specific terms in search engines. To address this risk, journalists might verify data results via their contacts and by providing context.

4. Misuse of data. This includes exposing or trading private data from data sets that should be aggregated and de-identified. Mecklin (2009) cites an example of potential abuse issues (privacy etc.) arising by United States government intentions to mine multiple databases to identify signals of terror activities. In the era of increasing regulatory transparency and interaction between consumers and journalists, 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)
5. News provider Demand Media creates stories based upon the popularity of search engine terms. The use of algorithms in selecting news topics may lead to popularity-driven news as opposed to important news that may not fetch as many “eyeballs”. For the purpose of this study, computational journalism does not solely include reliance upon the use of content algorithms for news items. This approach has been dismissed by analysts (Leonhard, 2009; Shirky, 2009; Jarvis, 2009) as “fast food” news with low content value. Reliance on content algorithms may damage the credibility of quality news brands.
6. Early adopter issues. Some software is still difficult and costly—e.g. transcription software is difficult to use and expensive with unreliable results.
7. There may be cultural differences between the working styles and professional priorities of investigative reporters and software programmers.
8. Role clarity. Clear delineation of roles between software developers, technology staff and journalists may prevent confusion over responsibilities and timelines. This is further complicated when projects include crowd-sourcing. Stenger comments on how issues arise when roles are unclear, and argues the role of technology staff generally is to maintain infrastructure:

they’re not necessarily there to deal with data and to generate insight about vast amounts of data. It’s a completely different skill set . . . So it seems like the division of labor question is being misaddressed by news organisations across the board—that IT and maintaining infrastructure is different than dealing with and processing news as data, especially for the purpose of getting insight out of it . . . If you’ve got a journalist who is data-literate, then the division of labor with IT smoothes out a little bit and the productivity goes up. (Stenger, 2008, para. 4)

Computational journalism does not automate the core tasks of journalistic work or enable the reduction of the number of journalists in newsrooms. Journalists “smoke out the most difficult-to-report situations . . . test glib assertions against the facts . . . probe for the carefully contrived hoax” (Steiger, 2009, para. 25).

## Summary

The forms of journalism are rapidly changing, but demand for journalists to inform the public and hold government accountable remains the same (Villano, 2009; Steiger, 2009). Investigative journalism is critical to the maintenance of democracy (Knight Commission, 2009; Murdoch, 2009) and

without investigative reportage, an array of major local and regional problems and corruptions would go unrevealed and unaddressed. 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. (Mecklin, 2009, p. 4)

Investigative journalism is under pressure from tightening financial margins and is relatively a high-cost and slow process. Computational journalism covers a variety of techniques and methods that, when successfully implemented, may reduce costs and increase the speed of investigative research. Correctly used, computational techniques provide a solid factual basis for news, from which journalists can verify results using human sources and provide context, then communicate the news appropriately using a variety of multimedia formats. This is important because:

while readership numbers for traditional print media are rapidly falling, the potential audience for online and computer-mediated news is soaring, which poses new opportunities for computation-savvy journalists. (Georgia Tech, 2008, p. 2)

Results from the techniques of computational journalism must be verified, qualified and explained, and this requires input from journalists. The use of computational journalism frees journalists from the low-level work of discovering and obtaining facts to verifying, explaining and communicating them.

## Further research

Further research may include:

1. A case study that applies computational journalism methods in an Australian, mass market context may demonstrate its applicability to news operations. Such a study might also investigate the impact on investigative journalism of emerging techniques and tools such as digital storytelling, multimedia reporting and co-creation.
2. An “anatomy” of computational journalism sources may be used to address the challenges of computational journalism. A second step may be to match interrogation software to specific journalism sources—i.e. identify the best computational journalism software for interrogating each source. This information, if codified, could form an important template for undertaking investigative journalism in a company, and should increase the speed and minimise any tedium or frustration of such research.
3. The attitudes of journalists and consumers towards computational journalism and co-creation may escalate or hinder takeup. Rodrigues (2009, para. 29) argues “in general, mainstream media outlets are still cautious in incorporating multimedia content produced by the public or this is not really part of their agendas”. A greater understanding of the basis for such attitudes may allow news providers to address concerns.

## References

- Andersen, M. (2009). Four crowdsourcing lessons from the Guardian’s (spectacular) expenses-scandal experiment. *Nieman Journalism Lab*, June 23. Retrieved May 10, 2010, from <http://www.niemanlab.org/2009/06/four-crowdsourcing-lessons-from-the-guardians-spectacular-expenses-scandal-experiment/>
- The Changing Newsroom. (2008). *Project for excellence in journalism*. Retrieved April 24, 2009, from <http://www.journalism.org/node/11961>

- Bruns, A. (2008), *Blogs, Wikipedia, Second Life and beyond: From production to produsage*. New York: Peter Lang.
- Daniel A., Flew T. & Spurgeon C. (2009). *User behaviours and intentions in digital media in Australia*. Paper presented at the Communications Policy & Research Forum, November 19-20, 2009, University of Technology, Sydney.
- Dedman, B. (2008). The color of money. *Power Reporting*. Retrieved December 2, 2009, from <http://powerreporting.com/color/>
- DigiDave. (2009, April 8). The rhetoric of journalism: Defining and re-defining what we do. Retrieved August 14, 2009, from <http://www.digidave.org/2009/04/the-rhetoric-of-journalism-defining-and-re-defining-what-we-do.html>
- Doctors Hangout Team. (2010). *Using iPhone app to crowd-source asthma research*. Retrieved January 11, 2010, from <http://www.doctorshangout.com/forum/topics/using-iphone-app-to>
- Doig, S. (2008). Reporting with the tools of social science. *Nieman Reports, Spring*. Retrieved January 2, 2010, from <http://www.nieman.harvard.edu/reportsitem.aspx?id=100075>
- Downie, L. & Shudson, M. (2009). The reconstruction of American journalism. *Columbia Journalism Review*. Retrieved December 14, 2009 [http://www.cjr.org/reconstruction/the\\_reconstruction\\_of\\_american.php](http://www.cjr.org/reconstruction/the_reconstruction_of_american.php)
- Georgia Tech. (2008). Journalism 3G: The future of technology in the field. *Symposium on Computation + Journalism, Georgia Tech. College of Computing, 22-23 February 2008*. Retrieved August 20, 2009, from [http://www.cc.gatech.edu/events/cnj-symposium/CJ\\_Symposium\\_Report.pdf](http://www.cc.gatech.edu/events/cnj-symposium/CJ_Symposium_Report.pdf)
- Google. (2010). *Google Living Stories*. Retrieved January 11, 2010 from <http://livingstories.googlelabs.com/>
- Greenslade, R. (2009, December 15). The sage lacks wisdom on newspaper business models. *Guardian*. Retrieved December 15, 2009, from <http://www.guardian.co.uk/media/greenslade/2009/dec/15/warrenbuffett-us-press-publishing>
- Hamilton, J. (2009). Tracking toxics when the data are polluted. *Nieman Reports, Spring*. Retrieved November 3, 2009, from <http://www.nieman.harvard.edu/reportsitem.aspx?id=100933>
- Hangarter, R. (2007, December 17). What is the recommender industry? *MSearchGroove*. Retrieved January 20, 2010, from <http://www.msearchgroove.com/2007/12/17/guest-column-what-is-the-recommender-industry>
- Howard, R. (2010, January 4). Five predictions on collaborative computing. *Social Computing Journal*. Retrieved January 20, 2010, from <http://socialcomputingjournal.com/viewcolumn.cfm?colid=871>
- IRE. (2009). *Philip Meyer Journalism Award: Investigative reporters and editors*. Retrieved November 30, 2009, from <http://www.ire.org/resourcecenter/contest/meyeraward.html>
- Jarvis, J. (2009, December 14). Content farms v. curating farmers. *Buzzmachine*.

Retrieved December 14, 2009, from <http://www.buzzmachine.com/2009/12/14/content-farms-v-curating-farmers/>

Knight Commission. (2009). *Information needs of communities in a democracy*. Retrieved October 30, 2009, from <https://secure.nmmstream.net/anon.newmediamill/aspen/kcfinalenglishbookweb.pdf>

Leonhard, G. (2009, December 13). Will algorithms run run our digital lives? *Media Futurist*. Retrieved December 14, 2009, from <http://www.mediafuturist.com/2009/12/will-algorithms-run-run-our-lives.html>

GovHack. (2009). *LobbyLens*. Retrieved December 12, 2009, from <http://team7.govhack.net.tmp.anchor.net.au/networkgraph.php>

Mecklin, J. (2009, January 10). Deep throat meets data mining. *Miller-McCune*. Retrieved August 20, 2009, from <http://www.miller-mccune.com/media/deep-throat-meets-data-mining-875>

Media Entertainment and Arts Alliance. (2008). *Life in the Clickstream: The future of journalism*. Retrieved January 20, 2010, from [http://www.alliance.org.au/documents/foj\\_report\\_final.pdf](http://www.alliance.org.au/documents/foj_report_final.pdf)

Mediastorm. (2010). *Washington Post*. Retrieved January 11, 2010, from <http://mediastorm.org/>

Meeker, M. (2009). Economy + Internet Trends. *Morgan Stanley Web 2.0 Summit, San Francisco, October 20*. Retrieved from [http://www.morganstanley.com/institutional/techresearch/pdfs/MS\\_Economy\\_Internet\\_Trends\\_102009\\_FINAL.pdf](http://www.morganstanley.com/institutional/techresearch/pdfs/MS_Economy_Internet_Trends_102009_FINAL.pdf)

Ministry of Justice. (2009). *The knowing or reckless misuse of personal data introducing custodial sentences: Consultation Paper CP22/09*. London. Retrieved January 11, 2010, from <http://www.justice.gov.uk/consultations/docs/data-misuse-increased-penalties.pdf>

Murdoch, R. (2009, December 1). There's no such thing as a free story. *Guardian*. Retrieved December 8, 2009, from <http://www.guardian.co.uk/media/2009/dec/01/ruPERT-murdoch-no-free-news>

Myers, S. (2009, April 14). Using data visualization as a reporting tool can reveal story's shape. *Poynter Online*. Retrieved July 10, 2009, from <http://www.poynter.org/column.asp?id=101&aid=161675>

Novak, J. & Wurst, M. (2003). Supporting communities of practice through personalisation and collaborative structuring based on capturing implicit knowledge. In *Proceedings of I-KNOW '03 Graz, Austria, July 2-4*. Retrieved January 20, 2010, from <http://www.know-center.tugraz.at/previous/i-know03/papers/cop/Novak COP.pdf>

Paynter, B. (2010, January 8). Global warming for deniers' fast company. *Fast Company*. Retrieved January 11, 2010, from <http://www.fastcompany.com/blog/ben-paynter/ben-paynter/global-warming-deniers>

Penberthy, D. (2010, January 13). Too busy holding meetings to do any actual work. *The Punch*. Retrieved January 20, 2010, from <http://www.thepunch.com.au/articles/too-busy-holding-meetings-to-do-any-actual-work/>

- Rodrigues, I. (2009, December 10). Making news with digital stories. *HASTAC*. Retrieved January 4, 2010, from <http://www.hastac.org/blogs/inezrod/making-news-digital-stories>
- Roth, D. (2009, October 19). The answer factory: Demand media and the fast, disposable, and profitable as Hell media model. *Wired*. Retrieved December 2, 2009, from [http://www.wired.com/magazine/2009/10/ff\\_demandmedia/all/1](http://www.wired.com/magazine/2009/10/ff_demandmedia/all/1)
- Shirky, C. (2009). *A speculative post on the idea of algorithmic authority*. Retrieved December 2, 2009, from <http://www.shirky.com/weblog/2009/11/a-speculative-post-on-the-idea-of-algorithmic-authority/comment-page-1/>
- Sifry, M. (2009). A see-through society: How the Web is opening up our democracy. *Columbia Journalism Review, January–February*. Retrieved July 12, 2009, from [http://www.cjr.org/feature/a\\_see-through\\_society.php](http://www.cjr.org/feature/a_see-through_society.php)
- Singel, R. (2009a, December 8). White House orders agencies to open up. *Wired*. Retrieved from <http://www.wired.com/epicenter/2009/12/white-house-orders-agencies-to-open-up>
- Singel, R. (2009b, November 30). AOL becoming automated, on-demand content factory. *Wired*. Retrieved December 14, 2009, from <http://www.wired.com/epicenter/2009/11/aol-automatic-content/>
- Steiger, P. (2009, October 14). Investigative reporting in the Web era. *McKinsey*. Retrieved December 1, 2009, from <http://whatmatters.mckinseydigital.com/internet/investigative-reporting-in-the-web-era>
- Stenger, B. (2008). Wiring Journalism 2.0. *Columbia Journalism Review, February*. Retrieved October 20, 2009, from [http://www.cjr.org/the\\_observatory/wiring\\_journalism\\_20\\_1.php](http://www.cjr.org/the_observatory/wiring_journalism_20_1.php)
- van Cuilenburg, J., Kleinnijenhuis, J. & de Ridder, J. (1988). Artificial intelligence and content analysis. *Quality and Quantity*, 22(1), pp. 65-97.
- Villano, M. (2009, June 8). Can computer nerds save journalism? *Time*. Retrieved June 10, 2009, from <http://www.time.com/time/business/article/0,8599,1902202,00.html>
- Wikipedia. (2010). *Computational journalism*. Retrieved December 20, 2009, from [http://en.wikipedia.org/wiki/Computational\\_journalism](http://en.wikipedia.org/wiki/Computational_journalism)

## Appendix 1      Examples of computational journalism

### **The Sunlight Foundation** (<http://www.sunlightfoundation.com/>)

The foundation has founded or funded projects aimed at revealing “the interplay of money, lobbying, influence and government in Washington in ways never before possible”. Bill Allison, a senior fellow at the Sunlight Foundation and a veteran investigative reporter and editor, summarises the non-profit’s aim as “one-click” government transparency, to be achieved by funding online technology that does some of what investigative reporters always have done—gather records and cross-check them against one another, in hopes of finding signs or patterns of problems.

### **EveryBlock** (<http://www.everyblock.com/>)

If you live in one of the eleven American cities EveryBlock covers, this site provides civic information (e.g. building permits and police reports), news reports, blog items and other Web-based information, such as consumer reviews and photos, all connected to an immediate geographic neighbourhood. Future plans include algorithms to take information from EveryBlock and other database inputs and write articles personalised to a neighbourhood and a person’s interests—e.g. a weekly news piece about crime in a neighbourhood and whether it has increased or decreased in relation to a month or a year ago (Mecklin, 2009).

*The following award winning news items used computational techniques for information extraction and visualisation, combined with strong investigative journalism:*

### **Guardian MP expenses**

*The Guardian* initiated an investigation into the expenses of Members of Parliament in the United Kingdom. They made available data (in Google spreadsheets) that consumers could search by member, constituency or item and then send comments or questions to *Guardian* staff about the data. *Guardian* journalists then investigated any questionable claims and built stories around them. A key benefit was that no other media entity was able or prepared to replicate the study. *The Guardian* owned the story because no other news source could easily replicate a similar study, and it provided the news entity with some degree of competitive advantage. It perhaps also enhanced the reputation of *The Guardian* for investigative journalism, and consumers were rewarded as well, with the MP expenses page noting who had reviewed most documents, and who had found the most interesting items. Over 20,000 consumers participated (with a participation rate of 56%) and 170,000 documents were reviewed in the first 80 hours. The flow of news items led to a government inquiry that found many members had incorrectly made claims. It allegedly took a software developer one week to build, with help from others in *The Guardian*, and an additional £50 to rent temporary servers (Andersen 2009).

The spreadsheet can be retrieved from:

<http://spreadsheets.google.com/ccc?key=phNtm3LmDZEObQ2itmSqHIA>

Expenses can be searched at: <http://ouseful.open.ac.uk/mpExpensesSearch.html>

An example of the results by MP can be retrieved from: <http://mps-expenses.guardian.co.uk/liberal-democrat/colin-breed/>

*The Guardian* MP Expenses news can be retrieve from:

<http://www.guardian.co.uk/politics/mps-expenses>

## **Color of money** (<http://www.colorofmoney.org/>)

This project grew from investigations into race and financial interests. It involved overlaying maps of the locations of middle income neighbourhoods in Atlanta (USA) where high African American populations were based, with locations of where banks rarely lend. It “disclosed that Atlanta’s banks and savings and loan institutions, although they had made loans for years in even the poorest white neighborhoods of Atlanta, did not lend in middle-class or more affluent black neighborhoods” (Dedman, 2008, para. 1). Consequently the *Federal Home Mortgage Disclosure Act* was expanded to provide more information to the public on the pattern of activity by all mortgage lenders.

Another 1998 investigation overlaid federal political campaign donations with census data to highlight race and ethnicity themes. It found that donations were predominantly from white wealthy communities, and this influenced government policy decisions. The investigative journalist responsible for this investigation (Bill Dedman) won a Pulitzer Prize for his work (Dedman, 2008).

## **Toxic waste**

James Hamilton mined the U.S. Environmental Protection Agency’s *Toxics Release Inventory* database, which contains information on chemical emissions by industry and the government (Hamilton, 2009). He used statistical and mathematical methods to tease out possible inaccuracies in companies’ reports of toxic releases. For example, his programs searched for violations of Benford’s law, the mathematical rule holding (accurately but counter-intuitively) that in many lists of numbers from the real world, the first digit will be “one” about 30 per cent of the time and “nine” less than 5 per cent of the time. Another of the programs looked for toxic release reports that were the same from year to year, on the theory that it is extremely unlikely for any industrial plant to emit precisely the same amount of toxins many years running. “That’s sort of the supplement (for investigative reporting), where you see the pattern”, Hamilton says. “It’s like a virtual tip . . . The whole idea is that we would be doing research and development in a scalable, open-source way” (Mecklin, 2009, pp. 3-4).

## **Hurricane damage**

Journalist Steve Doig layered wind speed data from Hurricane Andrew over millions of records of property data and damage reports. He discovered an undeniable pattern—the most heavily damaged areas were not those that experienced the highest winds. His investigation and reporting exposed poor building practices and standards, and a pattern of public corruption. The report titled ‘What went wrong’ was published three months after the hurricane (Doig, 2008).

## **Sudden Infant Death** (“Saving babies: Exposing Sudden Infant Death”)

Scripps Howard News Service reporters exposed bureaucratic lapses that hindered the search for causes of Sudden Infant Death. Using statistical tools, the team analysed the sharp differences in cause-of-death diagnoses among the states and produced the first rigorous proof of the value of the local and state child death review boards that only some jurisdictions use. A few months after the project ran, the then Senator Barack Obama introduced national legislation that would require medical examiners to make death-scene investigations in all cases of unexpected infant death (IRE, 2009).

**Car safety** (“Fatal failures” in the *Kansas City Star*)

Reporters analysed 1.9 million records from the National Highway Traffic Safety Administration to uncover NHTSA’s failure to consider non-deploying airbags as being a significant safety issue. The work by Casey and Montgomery suggested that nearly 300 people were killed each year in accidents when airbags didn’t inflate when they should have. Initially, NHTSA strongly disputed the findings, but finally did its own analysis and came to the same conclusions (IRE, 2009).

**US suburban policing** (“Too tough: Tactics in suburban policing” in the *Philadelphia Inquirer*)

Reporters studied arrest and court data from police departments in the suburbs that surround Philadelphia and found towns where blacks were being arrested in extraordinary numbers for minor offenses like loitering or jaywalking. Their follow-up reporting uncovered jails where thousands of illegal strip searches were being done, police dogs being used to control black children walking home from school, and traffic citations that were filled out in advance of arrests (IRE, 2009).

## **Appendix 2      Examples of computational journalism tools**

Government initiatives to increase transparency by enabling access to government datasets include:

- Australian Government—2.0 taskforce and mashups (<http://gov2.net.au/>);
- United Kingdom—Show us a better way (<http://www.showusabetterway.co.uk/call/data.html#gazette>); and
- The United States—Apps for democracy (<http://www.appsfordemocracy.org/>).

Online tools for analysis and mapping include Socrata, CKAN, Infochimps and Trendrr.

## **Appendix 3      Glossary**

### **3-D journalism**

The tools for converting lists and lines of numbers into beautiful, compelling images get more powerful every day, enabling a new kind of 3-D journalism—dynamic and data-driven. And in many cases, news consumers can manipulate the resulting image or chart, drilling into its layers of information to follow their own interests (Silfry, 2009).

### **APIs**

Applications programming interfaces (APIs) are data-aggregation tools that can process information in ways that reveal relationships and connections that otherwise might be obscured by the sheer volume of data at hand (Georgia Tech, 2008).

### **Citizen journalism**

When the people formerly known as the audience employ media tools in their possession to inform one another (Rosen cited in Digidave 2009).

### **Citizen reporting**

Citizens are often eyewitnesses not journalists and may record and make available a recording of an event, but not undertake activities to verify, make sense of the events and create news, e.g. the photos taken by mobile phones of the London train bombings or terror attacks in India.

### **CMS**

Content management systems

### **Computer assisted reporting**

This could be described as a historical version of computational journalism, before the rapid escalation of information, software and technologies and the digital economy.

### **Computational journalism**

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 (Hamilton & Turner, 2009).

### **Crowd sourcing**

Alternately called “network journalism”, crowd sourcing aims to organise groups of people through the internet to work on a single news item. Networked journalism rests its fate on two principles: the “wisdom of crowds”—the idea that collectives are more intelligent than individuals—and “distributed reporting”—the art of organising an online workflow, so that volunteers are efficient and happy to donate time to commit acts of journalism that in aggregate helps produce news. In distributed reporting, the work load is spread out (Digidave, 2009)—e.g. the analysis of MP expenses undertaken by *The Guardian*.

**Database journalism**

Similar concept to computational journalism.

**Digital storytelling**

Digital storytelling refers to using new digital tools to help citizens to tell their own real-life stories.

**Investigative reporting**

Investigative journalism is a form of journalism involving deep investigation of a single topic of interest. An investigative journalist may spend months or years researching and preparing a news report, which often takes the form of an exposé.