## Computational Journalism

<table>
<thead>
<tr>
<th>Journal:</th>
<th><em>The International Encyclopedia of Journalism Studies</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>IEJS0046</td>
</tr>
<tr>
<td>Wiley - Manuscript type:</td>
<td>Entry</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>11-Dec-2017</td>
</tr>
<tr>
<td>Complete List of Authors:</td>
<td>Anderson, C.W; University of Leeds, Media and Communication Caswell, David; Structured Stories</td>
</tr>
<tr>
<td>Keywords:</td>
<td>computational journalism, structured journalism, algorithms, data journalism, epistemology</td>
</tr>
</tbody>
</table>

**Abstract:**

Computational Journalism is defined as a practice in which journalistic knowledge is represented computationally, as systems of discrete categories or as numbers, during reporting, analysis, distribution or consumption. This is in contrast to traditional journalism, in which journalistic knowledge is reported, analyzed, distributed, and consumed as natural language text or speech, whether in analog or digital media. Journalistic knowledge is defined as externally-represented knowledge that is under human editorial control in the service of journalistic values. In this entry, we review the development of computational journalism as a practice and as a field of study before examining the inherent tensions which structure it as a field.
Computational Journalism

Abstract

Computational Journalism is defined as a practice in which journalistic knowledge is represented computationally, as systems of discrete categories or as numbers, during reporting, analysis, distribution or consumption. This is in contrast to traditional journalism, in which journalistic knowledge is reported, analyzed, distributed, and consumed as natural language text or speech, whether in analog or digital media. Journalistic knowledge is defined as externally-represented knowledge that is under human editorial control in the service of journalistic values.

In this article, we review the development of computational journalism as a practice and as a field of study before examining the inherent tensions which structure it as a field.

Keywords

Computational Journalism, algorithmic journalism, quantitative journalism, data journalism, automated journalism, structured journalism

Computational journalism is an emerging area of journalistic practice, as well as a somewhat diffuse category of academic scholarship. Like many topics of journalism studies, it is as much defined by journalistic work as it is by academic research. In this entry we first attempt to clarify what exactly this work is, and how it has evolved. We next look briefly at the emerging academic work on this topic, pointing out how it has largely avoided high level theorizing and has focused on empirical research, much of it qualitative. We conclude by pointing out some of the most important tensions in discussions of computational journalism as well as likely avenues of future research.

This entry necessarily requires some definitional foundations, despite the lack of a clear consensus perspective in the field. We have done our best to minimize the impact of these definitional choices, and to focus instead on the substance of the topic.

History

Computational forms of journalism, like the term ‘computer’ itself, pre-date the development of computing machines. Early examples of the use of numerical or categorical expressions to communicate journalistic knowledge can be found in the very first edition of The Manchester Guardian, published on May 5th, 1821, and in the December 22nd, 1848 edition of The New York Tribune. More broadly, pre-digital journalism was informed and shaped by structured information schemas maintained by newsrooms in the form of systems of cross-indexed cards based on taxonomies and categories of people, entities, concepts, events and locations. These products and systems were produced manually by newsroom employees for whom the term ‘computer’ referred to a job description rather than to a technological device. Although such manual categorization and computation was not a significant factor in journalism prior to the availability of digital computing, its existence illustrates that computational journalism is fundamentally an interpretation of the nature of journalistic knowledge, rather than merely a technological practice. It is this interpretation, rooted in specificity and abstraction reminiscent of the scientific method, which forms the cultural basis of computational journalism.
The automation of computation using digital computing machines began in the 1940’s, and digital computers soon provided a technical infrastructure from which new machine-aided forms of computational journalism could emerge. In 1952 a digital computer was used by CBS News to predict the outcome of a presidential election using partial results, and by the 1960’s news stories originating from computational analysis were being published by the Detroit Free Press (survey records analyzed by Phil Meyer) and by the Miami Herald (government records analyzed by Clarence Jones) – treatments of journalistic knowledge that would be practically impossible using manual computation. As digital computing became increasingly ubiquitous in society and in newsrooms during the late 1980’s, a variety of computational practices for gathering or analyzing journalistic knowledge emerged ad hoc. These practices were typically pioneered by individual journalists experimenting with computational techniques and with computational representations of journalistic knowledge, for particular objectives on particular stories. They included computer-aided analysis of public records, spatial analysis of digital maps, and the assembly of small datasets to enable the detection of patterns – an approach exemplified by The Boston Globe’s Spotlight team’s pivotal use of a spreadsheet in reporting their Pulitzer prize-winning 2002 story about the Boston Archdiocese’s cover-up of child abuse. Collectively these techniques were gathered under the generalist term ‘computer-assisted reporting’ (CAR), and became increasingly formalized and institutionalized through the work of the National Institute for Computer Assisted Reporting (NICAR) in the United States. Although much-maligned, the term CAR accurately reflects a central characteristic of this era of computational journalism - that its techniques and representations were developed and applied by working journalists who adopted computational methods to solve specific practical problems within the framework of traditional journalism.

The emergence of the internet in the mid-1990’s, the digital distribution of news and the associated economic disruption for news organizations, together led to a substantial change in the status of computation within journalism. Instead of being perceived as merely an exotic interpretation of journalism or as a collection of useful tools, computational journalism was increasingly seen as a possible means by which journalism could maintain its relevance and economic sustainability within an unfamiliar networked information ecosystem. This newfound status professionalized computational journalism within large newsrooms, as reflected by new job titles and department names, increased budgets for ‘digital’, and by the appearance of trained computer science professionals engaged in prioritized multi-person software development projects within newsrooms.

By the late 2000’s new areas of specialization began to emerge, including data journalism (the use of data in reporting, analyzing and writing or visualizing stories), automated journalism (the automatic writing of articles from data), sensor journalism (the automated reporting of news from electronic sensors), structured journalism (the manual reporting of news as data) and, more recently, conversational journalism (the communication of news via conversational interfaces, such as chat bots). Each of these variations emerged relatively independently and their domains overlap substantially, however each is centered on an interpretation of journalistic knowledge as data and each is therefore fundamentally computational. Although aligned with traditional language-centered journalism and its values, and although still highly experimental, these new natively-computational forms represent a sustained and intentional effort to establish not only computational interpretations of journalistic knowledge, but computational interpretations of journalism itself. In other words, they seek to explore new systems of journalism that might eventually operate alongside traditional language-centered journalism, rather than merely within it. The possibility of such a development has been made somewhat credible by the speed with which the digital distribution of news was assumed by search and social media technology firms.
The considerable attention given to these revolutionary variants of computational journalism has somewhat overshadowed the simultaneous evolution of computer-aided reporting tools. Unlike the story-specific projects of technology-savvy journalists that typified 20th Century CAR, a new generation of general-purpose computational tools has emerged from significant development efforts conducted by small teams of experienced software developers supported by foundation funding. These projects included Document Cloud, The Overview Project and Investigative Dashboard, all of which seek to facilitate the exploitation of large repositories of text documents for investigative journalism. The vision behind these projects in large part originated at an influential 2009 workshop titled “Developing the Field of Computational Journalism”, which focused on computational tools to aid investigative journalists in finding, analyzing and reporting stories. Although document-centric and unambiguously aligned with traditional language-based journalism, these projects nonetheless include significant computational representations, especially named entities and other elements of natural language processing. While they do not seek to reimagining what journalism might become in the digital age, their day-to-day use in newsrooms may be more impactful in the short term than the more exotic ‘x-journalism’ variants.

The growing status of computational journalism within newsrooms has been paralleled by the appearance of similar experimental initiatives originating outside of journalism and organized as commercial technology start-ups, but seeking to apply computational practices to journalistic ends. These projects were founded by entrepreneurs, computer engineers and hackers attracted to challenges and opportunities associated with reimagining journalism for the digital environment, and often employed journalists in editorial roles. Prominent examples of this variant of computational journalism were Circa (a structured journalism start-up founded in 2012 by three technology entrepreneurs), Timeline (a contextual news start-up founded in 2015 by a technology investor) and InkaBinka (an AI-driven news summarization app founded in 2014 by a technology entrepreneur).

A coherent identify for computational journalism as a well-defined field has not yet emerged. The appearance of such an identity may be close at hand, however. Scholarly interest in typologies of computational practices in journalism has increased, while the appearance of increasingly integrated computational journalism systems and organizations in the mid-2010’s suggests that the field may define itself, de facto, before it is defined by scholars. Examples of such integration include The Los Angeles Times’ Quakebot, which combines sensor journalism and automated journalism and which is operated by self-described data journalists, and Quartz’ s Bot Studio, which develops conversational journalism interfaces using editorial content assembled as structured journalism. This blurring of distinctions between the variants within computational journalism is occurring while references to ‘artificial intelligence ’ (A.I.) in journalism appear frequently in industry publications, often in reference to automated journalism or conversational journalism ( ‘news bots’). While much of this rhetoric is sensationalist speculation, or a naïve interpretation of current technologies, or both, there is also a sense that the portfolio of increasingly capable and integrated computational techniques being applied to journalism may be adding up to something greater than the sum of its parts. The recent appearance of hugely successful AI-driven next-generation news aggregation platforms in Asia, such as SmartNews and especially Toutiao, suggest that computational approaches to journalism are more relevant than ever. While true journalistic ‘A.I.’ is unlikely to emerge in the foreseeable future, it should be remembered that structural changes in information environments, such as those occasioned by the printing press or by the internet, can have far-reaching and unanticipated consequences. The emergence of a successful and widely-used computational journalism environment, were that to occur, might constitute such a change.
Computational journalism, like journalism generally and like much of computer science, is an applied practice rather than a theoretical discipline. The early literature of the field was similarly practical. Phil Meyer’s influential 1973 book “Precision Journalism” and its revisions, often considered the founding manifesto of computer-assisted reporting, was essentially a how-to manual for applying field, statistical and computational practices from the social sciences to journalism. Early treatments of computational journalism in academic journals typically consisted of descriptions of computer use by pioneering newsrooms. The practicality of this early literature reflected the then common view of computational journalism as a peripheral technique, supportive of traditional journalism, to be considered, explored and applied where it could be useful.

By the early 2010’s, as computational journalism began to be perceived as possibly more than merely an adjunct to traditional journalism, analytical scholarship of computational journalism began to appear. This work had greater urgency than early work, and typically sought to understand important nuances and repercussions associated with particular forms of computational journalism, particularly data journalism. Papers from this analytical literature consist of case studies, ethnographic observations, studies of content or algorithms, and surveys or interviews of journalists, often accompanied by attempts to more crisply define particular forms of computational journalism. This largely qualitative and descriptive research may avoid high level theorizing or an elaborate methodological apparatus, but it generally provides a critical analyses of its subject matter and often raise practical concerns about possible consequences. Some particularly influential examples of this genre of the computational journalism literature include Lewis and Usher (2013) and Parasie and Dagiral (2013).

The increasing sense among journalism scholars, and others, that the impact of computation on journalism was likely to be significant, combined with proliferating and ambiguously defined applications of computation in journalism, has led to proposals of frameworks that seek to describe computational journalism as a complete and coherent field. They generally propose a set of dimensions or lenses with which to organize and understand the field, and then interpret examples of computation in journalism in terms of those dimensions or lenses. The perspectives embodied in these frameworks include a craftsmanship view, a typological view (Coddington’s 2015 application of professionalism, openness, epistemology and ‘vision of public’ dimensions), a sociological view, a platform view (Diakopoulos’s 2016 ‘tooling’ perspective), and a historical view (Anderson’s 2018 history of data journalism in the United States). These early theories are typically taxonometric rather than developed from first principles. While it is clear that a distinct field exists, it has originated as a collection of ad hoc responses to opportunities presented to journalism by the appearance of computational tools. This is similar to the practice of computer science, which much more focused on utility than on theoretical foundations, however unlike computer science computational journalism still lacks those theoretical foundational describing what it is and what its atomic processes and components are. Journalism, as a pure practice, has avoided the need for such theoretical foundations, however it is not clear that computational journalism, as a hybrid of craft and science, can do likewise. Although computational journalism has made considerable progress without such a theory, its status as an ad hoc collection of tools and tasks may thwart the establishment of a coherent and strategic research agenda.

These utilitarian, analytical and theoretical literatures associated with computational journalism are primarily sociological rather than technical. They generally seek to understand, critique and define their subject, rather than to participate in it on its own terms.
or to advance it as a technological practice. A separate literature has appeared seeking to fill that gap, however it has originated not from journalism or journalism studies but from computer science. This community is centered on a handful of academic workshops associated with technical conferences in computational linguistics and knowledge engineering and focuses specifically on computational interpretations of news. A handful of technical papers relating to representing and organizing news has also appeared at several prominent computer science conferences, such as SIAM, and in prestigious technical publications such as IEEE. This computational journalism research community is interested primarily in technical experimentation, usually without immediate expectation of practical or commercial application, and its work is often based on long-standing foundational scientific and technical research in computer science. A flagship project associated with this literature is NewsReader - a EU-funded research effort active from 2013 to 2016 that sought to automatically extract structured news events and storylines from large corpuses of text articles. Although journalists and journalism scholars have generally not participated in this community, some computer scientists employed by newsrooms and some technical researchers affiliated with journalism schools have. A technical symposium series, founded by a computer scientist and held annually at universities experimenting with computational journalism programs has been particularly important in facilitating this cross-disciplinary communication. This literature has seen no known attempts to contextualize its work through the development of theoretical frameworks.

Tensions

Computational journalism is, obviously, an inherently multidisciplinary field with deep roots in computer science and in traditional journalism. It is also, less obviously, a field that encompasses multiple conceptual paradigms. Few practitioners or researchers working in computational journalism are fluent in each of these paradigms, and therefore the field has developed in the presence of tensions between its cultural constituents. The many multidisciplinary partnerships and interactions that together make up computational journalism as a coherent field owe much to traditional journalism’s search for sustainability and relevance in a digital information ecosystem, and also to the inability of software engineers to interpret and apply useful editorial judgment computationally. Many of these partnerships may therefore be of necessity rather than of choice.

Probably the most significant cultural boundary present at the heart of computational journalism is the boundary between the sciences and the humanities. Defining computational journalism by its representation of journalistic knowledge as structured data entails assumptions about the need for precision, the value of accessibility to logical and statistical inference, and a separation between the collection of data and its communication, which taken together align the field with the quantitative social sciences. In contrast, traditional journalism’s assumptions are rooted in the representation of journalistic knowledge as natural language, informed by nuance, ambiguity and style, and inextricably tied to individual authorship. These distinctions, and others, have proven to be formidable barriers to the development of, and adoption of, computational journalism systems. True partnerships between editorial and technical constituents are difficult to forge and are therefore rare, and many computational journalism projects fail due to insufficient appreciation of editorial realities during development, lackluster adoption of technologies in newsrooms, or both.

The practical challenges in establishing productive working relationships between technologists and journalists on computational projects are underpinned by a more substantial barrier between scientific and humanistic cultures, rooted in values rather than
merely in assumptions. This ‘values barrier’ in computational journalism is a skirmish in a larger contest as to whether the ultimate source of authority within post-modern society should lie with individual human judgment or with the externally validated data. Unlike the debate between the sciences and the humanities (which are basically differences rooted in assumptions that can be overcome with effort) differences rooted in values might present more of a challenge to progress in computational journalism. Taken to extremes this contest may even work to negate the field entirely by, on one hand, seeking data-centric ‘journalistic’ solutions that are devoid of human editorial judgment and, on the other, seeking a retreat into traditional text and accepting associated losses of sustainability and relevance.

Participants in computational journalism also exhibit differences related to their acceptance of abstraction. Computer science, and computational thinking itself, is essentially a practice of abstraction which seeks to define classes of problems and then apply generalized solutions to them. Traditional journalism, on the other hand, is a practice of specifics, in which each story is considered unique and in which solutions are sought in the form of improved tradecraft. This affects the design of computational systems for journalism because the very participants most qualified to define classes of problems within journalism (i.e. working journalists) are often uninterested in abandoning the nuance of specific stories and resist fitting them to a common mold. Meanwhile the participants most skilled at forming and working with domain-specific abstractions (i.e. practicing technologists) are unable to grasp the patterns and features of editorial subject matter. This cultural distinction between abstraction and specificity may extend into the scholarship of computational journalism, perhaps biasing the field towards ethnographic and descriptive studies, and away from theory building.

Finally, the trajectory of digital technology in recent years has also had an indisputable effect on the constituent components of computational journalism. The dominant and growing status of computation within society may endow technologists working in journalism with optimism, confidence and economic security, whereas the technology-driven decline in the economic stability and status of traditional journalism may engender very real concerns among journalists about deskilling, employment status and loss of control. This asymmetry in exuberance about technology and its potential may also have an impact on the exploration and adoption of computational methods in journalism.

Despite these substantial cultural differences, computer scientists and journalists are increasingly finding ways to work together productively, often within major news organizations and even within newsrooms - the Washington Post is a particularly prominent example. This cooperation is clearly driven by more than expediency, and reveals an underlying sensibility common to (some) journalistic and technological cultures - a muckraking, disruptive and skeptical suspicion of entrenched power. The prototypical narrative of programming generally, and of open source software in particular, is of independent, free-thinking individuals who, through their own ingenuity and resourcefulness, succeed in upending the powers-that-be and usher in a new, more truthful, reality to the benefit of society. This is also journalism’s prototypical narrative. This common self-description is explicitly expressed in the name of the primary grassroots organization promoting interaction between journalism and computation - Hacks and Hackers. Computational journalism, having gradually emerged as a sprawling, ad hoc collection of tools and techniques, riven by tensions, may be on the cusp of becoming a coherent field and a mainstream practice. If it succeeds it will be at least partially due to this shared ethos.
References


