Digital archives, big data and image-based culturomics for social impact assessment: Opportunities and challenges

Kate Sherren[a],[*], John R. Parkins[b], Michael Smit[c], Mona Holmlund[d], Yan Chen[a]

[a] School for Resource and Environmental Studies, Dalhousie University, Halifax, NS B3H 4R2, Canada
[b] Department of Resource Economics and Environmental Sociology, University of Alberta, Edmonton, AB T6G 2H1, Canada
[c] School of Information Management, Dalhousie University, Halifax, NS B3H 4R2, Canada
[d] Department of History, Dalhousie University, Halifax, NS B3H 4R2, Canada

ARTICLE INFO

ABSTRACT

Social impact assessment (SIA) is well-established but uses conventional approaches that have become less effective in recent decades, particularly in relation to declining survey response rates and a lack of youth engagement. Images from digital archives and social media sources are poised to advance the research and practice of SIA by transcending text-based methods with insights into changing landscapes, and human engagement with them. This viewpoint describes progress, challenges and cautions toward the development of such tools (defined as culturomics), using hydroelectricity cases to illustrate potential approaches. These tools build on foundational work in a range of disciplines, including the humanities and computer science. We describe necessary advances in machine learning, image digitization, and data aggregation and visualization techniques, as well as ways to ensure that such tools are carefully tested, applied and interpreted. Challenges include the automation, acquisition and management of datasets, and using these tools appropriately and equitably. Critically, culturomics of any kind must not be used as a replacement for engagement with people, but as complementary to inclusive stakeholder engagement.

1. The need for new tools for SIA

Environmental assessment is a long-established practice with wide-ranging procedures for data collection and analysis, focused on the anticipation of impacts from policy or project proposals. Within social impact assessment (SIA) in particular, conventional tools for assessment include local surveys, analysis of secondary data (e.g., census data), interviews with key informants, and a range of methods for anticipating social impacts such as the calculation of multiplier effects, straight line analysis, and comparative case studies (Asselin and Parkins, 2009). Although there are some efforts to update these methods with more sophisticated approaches to public participation (Sinclair and Diduck, 2017) or simulation models (Karami et al., 2017), many of these advances remain within the realm of academic work, with limited application in the practice of SIA. Although the SIA literature has traditionally focused on *ex ante* assessments of impacts, this focus is changing with growing attention to *ex post* assessment where long-term policy or project monitoring comes into focus. This shift in the definition of SIA is reflected in the work of Vanclay et al. (2015), where SIA is described as:

"the process of identifying and managing the social issues of project development, and includes the effective engagement of affected communities in participatory processes of identification, assessment and management of social impacts" (Vanclay et al., 2015, p. iv).

Although the identification and management of social issues is a point of focus in SIA literature, methods for doing this work have not advanced significantly in recent decades. SIA remains entrenched in conventional public engagement procedures and analysis using local household surveys and census data (Parkins and Mitchell, 2016). Such conventional methods are rendered increasingly ineffective because of steep decreases in survey response rates and associated growing biases in those who participate (Connelly et al., 2003; Stedman et al., 2016). Additionally, after a development is approved, follow-up social monitoring is rarely undertaken to confirm the predicted effects or to assess the effectiveness of impact mitigation.

A real weakness of SIA is the bias introduced by voluntary public participation, meaning younger demographics are often under-represented (Checkoway et al., 2005). Though apathy toward politics and public affairs has generally increased over the past 30 years, people under the age of 25 show the biggest decrease in interest (Delli Carpini,

2000). Younger generations have very different conceptions of civic engagement, but should not be ignored (Blandford et al., 2015). Much of the engagement around public issues that does occur within this demographic cohort – Generation Y or so-called Millennials who were born from the early 1980s to early 2000s – happens on social media where this generation documents their lifestyles and attitudes (Delli Carpini, 2000). Despite its increasing ubiquity, social media is generally only used in SIA to share information and occasionally monitor online sentiment (e.g. *Twitter*). Such uses are important for tracking opposition around specific projects but less nuanced when it comes to understanding the social and cultural implications of those proposals (Hanna et al., 2016).

In an increasingly digital world, practitioners of SIA have a growing opportunity to leverage a wider range of datasets, including digital images and associated text in archives and social media sources (Esteves et al., 2012). The term 'culturomics' encompasses this growing field of research, described as the use of text-based corpora, or collections, to understand culture. The inventors of this term leveraged the *Google Books* database to track the usage of words and phrases (and thus one lens on culture) over time (Michel et al., 2011). They liken culturomics to the microscope or telescope, as it makes visible for study a dimension that was previously largely hidden. Even using only the text associated with online activity, such as search engine queries and *Twitter* traffic, researchers have been able to track cultural trends and monitor environmental conditions (Di Minin et al., 2015; Ladle et al., 2016). Given the growth in image-dominated datasets, however, we are thinking very small indeed if we do not leverage the increasing volumes of online images, particularly those generated by citizens (Sherren et al., 2017b). Unlike predominantly text-based sources such *Twitter*, which is more explicitly polemical, messages carried by landscape images on social media often *imply* landscape perceptions, preferences, and lifestyles. Research using social media is typically nonreactive, as subjects are unaware of being studied. This may not mean that such research is immune to biases such as social desirability. Social desirability oriented toward peers and other expected audiences will undoubtedly bring different perspectives to SIA, however, than bias introduced by government or proponent intervention. Image collections that provide a longer historical perspective are also held in other increasingly digital databases, such as newspapers/sites and archives.

Advances in publishing technology have come with the widespread means to easily produce and publish image, audio, and video and therefore a need to widen the lens through which we observe human/environment issues. We focus here on the potential of leveraging online images and technical capabilities in image processing to improve SIA tools. We exclude here the understanding of environmental and social change that can be determined through analysis of aerial imagery such as from satellites, either using feature recognition or spectral analysis. Drawing on Vanclay's (2002) distinction between social impacts and social change variables, we see culturomics contributing to several social change domains. These include demographic processes (e.g., influx of tourists and seasonal workers, displacement and dispossession), and socio-cultural processes (e.g., perceived differences between community groups, changes in cultural values). Geographic processes are particularly relevant here, with attention to how social change accompanies land use conversion and diversification, urbanization, gentrification, changing transportation systems and other physical changes and uses of the landscape. Given the exponential growth in digital archives and the recent emergence of culturomics, in this commentary we describe the interdisciplinary state-of-the-art that is converging to enable new tools for SIA, using hydroelectricity as a case study, and describe some of the opportunities and challenges that accompany the development and use of such tools.

2. The interdisciplinary state-of-the-art

A 'pictorial' turn is afoot across many fields that is increasingly well-described (Bachmann-Medick, 2015) and already plumbed for insight in the arts, humanities and social sciences (e.g. Graham et al., 2011; Pink, 2003). It includes approaches to describing contemporary culture, as well as tracking cultural change through time. For the latter, we can build on approaches in art history that observe cultural shifts such as landscape perceptions and cultural norms (Halkes, 2006). More recently fields such as material and visual culture studies explore cultural changes in environmental behaviour through artefacts such as paintings, advertisements, television shows and everyday objects (Bennett and Joyce, 2010; Shove et al., 2007). These fields demonstrate that structured analysis of images can help us understand not only their contents and meaning, but also the emotional power that images acquire as they move between individuals and across cultures through various media (Belting, 2011). One can easily imagine what longitudinal insights could be established using increasingly digitized newspaper and archival image corpora.

The integration of social media into contemporary society provides emerging opportunities for describing today's culture. Images are being shared online with a frequency and density that is culturally unprecedented. These images reveal personal values, lifestyles, landscape uses and perceptions. Because of shifting technologies – which social media software is being used, its uptake and demographic biases, and its internal rules for use – tracking change over time is less possible with social media. Images hold great potential, however, for understanding how the connections between features and values are made over space. Existing methods for doing this, for example photo-elicitation or values mapping with research participants (Brown and Fagerholm, 2015; da Silva and Antunes, 2014), are conceptually rich but time-consuming in both data collection and processing. Through passive (i.e. non-reactive) data collection, social media offers a clear opportunity to increase sample size and compensate for biases introduced by researchers using more active approaches (i.e. those *generating* rather than collecting data). Barry (2013), for example, demonstrated such image-based approaches can help develop nuanced understandings of public concerns about grazing on public lands, and help avert conflict over land management issues.

In recent years the techniques for automation in the processing of such data volumes, whether social media or archival, have matured in their sophistication and application. 'Big Data' refers to the phenomenon of having large volumes of data in a variety of unstructured formats accumulating at high velocity. The human brain can readily comprehend and analyze images and unstructured text, but only in relatively small amounts. Computer scientists are developing sophisticated algorithms to help in cases where we have too much data and not enough humans. Machine learning algorithms can learn how humans processed a set of images and text, and apply that understanding on a vastly larger scale. Image recognition algorithms, once trained, can generate descriptions of the events, objects, and interactions occurring in an image (Fang et al., 2015; Karpathy and Fei-Fei, 2015). Automated approaches can also be used to determine where photographers sharing work online actually live, even if their profiles are anonymous (Rugna et al., 2012). Combining images and caption text within analyses may indeed help balance the varying weaknesses of each source, including the use of colloquial language and intentional misspellings (Agarwal and Yiliyasi, 2010). Technology companies often deploy these algorithms for real-world applications through cloud-based web services (e.g. IBM Bluemix Watson APIs, Google Cloud Machine Learning, Microsoft Azure Machine Learning Studio). Application of digital media analysis also extend to emerging areas of computational impact assessment and social computing, for instance to assess the individual and broader impacts of social change documentaries and projects (Diesner et al., 2014; Diesner and Rezapour, 2015; Rezapour and Diesner, 2017).

Leveraging images to track social, environmental and cultural trends builds on advances in other fields, sitting at the nexus of several new and established methodologies. As mentioned above these ideas are related to the proliferation of Big Data and ubiquitous public