# Between technical constraints and epistemic assumptions

## The socio-technical influences on making web-based visualizations a device of social analytics

Contact: Anders Koed Madsen, Department of Organization, Copenhagen Business School, akm.ioa@cbs.dk

**Abstract**

Digital traces and the visualizations they give rise to are increasingly used as a source of data with which to understand the social world. This paper presents an analysis of eight projects engaged in reconfiguring practices of social analytics on this basis. Despite being carried out in response to different problems it is shown that these projects face common challenges in relation to the inevitable distribution of data-formats to third party actors and the need to balance machine intelligence and human intuition. For each challenge the paper identifies two opposite ´digital choices´ made to meet them. These choices indicate a room of flexibility within the constraints posed by the challenges. The paper furthermore shows how digital choices are legitimized through different ´web-frames´ that establish a relationship between the choices and the context in which the visualizations are to be used. Through the conceptualization of these challenges, choices and frames the paper pinpoints central socio-technical influences that will shape the way web-based visualizations can potentially be used as an analytic device of policy-planning in the future.

# 1. INTRODUCTION

When Google began to use patterns in hyperlinks as the empirical basis for determining the relevance of web-pages at the end of the 1990´s they did not just redefine the practice of search (Brin and Page 1998). Their rapid success made a compelling case for the argument that traces left in the digital world can be used as a basis for handling the complex and rapidly growing body of information in today's society. The fact that hyperlinks were widely accepted as useful indications of information-relevance sparked a belief in the possibility of re-purposing other forms of digital traces as a basis for understanding other forms of social dynamics as well. Within academia this belief has been articulated in the call for ´computational social science´ (Lazer et al 2009) and it has recently been influential in managerial trends taking advantage of the rise of ´big data´ (McKinsey Global Institute 2011; Anderson 2008). Projects under these headings have been engaged in utilizing software-tools to harness digital traces such as links, tags and tweets and make them comprehensible through ´web-based visualizations´ in the form of, for instance, network graphs and tag-clouds.

Web-based visualizations are a tool of social analysis that differs from traditional tools such as surveys and focus groups. This means that they have the potential to reconfigure organizational attention, thinking and decision-making in a way that resembles the way Google´s approach to search has reconfigured information-filtering on the web (Battelle 2006). This is not least true within the field of policy-intelligence and development work (World Economic Forum 2012). But new technical affordances do not come with unambiguous guidelines for their use and new analytical devices are rarely entering the world in a ready-made and widely accepted form. They are rather born as underdetermined tools that need to be experimentally stabilized, made sense of and harmonized with existing social practices (Marres forthcoming). On the basis of this approach to emerging technologies the paper sets out to answer the following question.

*What are common challenges for project leaders that use web-based visualizations as analytic devices and what is the scope of variation in the way these challenges are handled and the way the proposed solutions are legitimized?*

This question is answered through interviews with eight project leaders that are experimenting with the use of web-based visualizations as well as qualitative analysis of documents they suggested as relevant to understand their work. The project leaders work in different sectors such as advertizing, social science, technology foresight, policy-planning and military intelligence. By identifying similar challenges across different sectors the paper identifies some central trade-offs involved in the use of web-based visualizations. Differences in the way these challenges are handled and differences in the way the proposed solutions are legitimized is, to the contrary, used to indicate different scenarios of use of this analytical method.

The paper is organized so that section 2 presents the basic concepts of ´digital choices´ and ´web-frames´ that guides the analysis. Section 3 explains the empirical methodology of the study. Section 4 identifies two technical challenges that are central to all project leaders. One is the necessity to distribute data-formatting to third party actors and the other is the need to balance machine intelligence and human intuition. The section ends by outlining the most different digital choices made to meet them. Section 5 identifies assumptions about data-benchmarks and conditions of decision-making as two central themes in the way such choices are legitimized through the construction of web-frames. It furthermore outlines the most different assumptions that go into constructing such frames across the cases. In section 6, the analytical results are used to argue that the final shape of a web-based visualization is dependent on a socio-technical chain of selection-mechanisms that involves technical constraints in getting the data as well as the influence of the epistemic frames in which this data is meant to travel. The potential of using visualizations as devices of policy-planning is used as an example to drive out potential policy-implications of these findings.

## 2. DIGITAL CHOICES & WEB-FRAMES

The concept of ´digital choices´ was originally introduced to strengthen the analytical focus on choices about the design, use and governance of the Internet (Dutton 2007). Examples are the design of cost-structures related to the production of digital content and technical choices about the way information is filtered. Such choices are interesting because they have the potential to reconfigure people's access to relevant information and the fact that these choices are ´digital´ means that they are influenced by technical affordances and constraints related to existing Internet-technologies. But it is at the same time emphasized that digital choices are just as much shaped by their entwinement with existing organizations and societies. In relation to the Internet it is, for instance, argued that digital inequalities, the popularity of paradigms that highlight the transformative potential of the web and expectations of web-users are shaping the central digital choices. The Internet can, accordingly, result in different kinds of access to information depending on the socio-technical chain that influences the digital choices that shapes its infrastructure.

Web-based visualizations are smaller object of study than the Internet but they have the potential to reconfigure the way people and organizations access information as well. They are here defined broadly as spatially organized depictions of information that are constructed by harnessing digital traces across the web. A well-known example of such a visualization is Facebook´s ´friend wheel´. If you use this application to visualize your social network you will get a depiction of your friends that rely on a specific form of structural information. It is based on traces of friendship ties left of Facebook´s interface and it is built to guide your attention towards friends that serve as connections between coherent groups of friends. It is important to note that this reconfiguration of your attention towards your friends will be based on a distributed set of digital choices (Madsen 2012). Examples of such choices are Facebook´s design of their interface, their infrastructural choices about which actions that are required to be categorized as friends and the assumptions about groups and bridges that are built into the visualization-software. These choices are shaping the way traces of friendship are turned into a visualization that provides them with new structural qualities and make you to think about them in a new way (Manovich 2008).

The digital choices used to build web-based visualizations are also not made in a technical vacuum, however. Just as the access to information provided by the Internet is shaped by existing institutions and social practices so is the attention-span provided by web-based visualizations. Keeping with the example above one can say that the operationalization of friendship need to have some sort of resonance in the culture in which it is used. This need for resonance is captured in the concept of a ´technological frame´ that highlights the importance of forming ideas about how technologies work and how their future use can be envisioned (Flor and Meyer 2008). In the context of the web-based visualizations such frames will be denoted as ´web-frames´ that present ideas about the problems these visualizations can solve and the way they can support the practice of social analysis. The web-frame aims at aligning digital choices in the construction of a visualization with central epistemic assumptions in the context in which it is to be used. In order to be successful it must integrate heterogeneous elements such as hardware, software codes, problem formulations and dreams of the future.

## 3. THE RESEARCH DESIGN

The projects-leaders interviewed are all using web-based visualizations as an analytical device to guide their attention to information about specific social dynamics. They are chosen on the basis of a ´most different´ case-study design that is suited to identify similar challenges across projects with interests in quite different social dynamics (Flyvbjerg 2004). These dynamics vary from cultural tensions around brands to innovation paths around emerging technologies. The dynamics of interest are listed in column two in table 1 below. The projects picked out for analysis was found by browsing presentations at relevant academic and business conferences as well as by following suggestions from the first interviewees and other experts in the field. This snowball technique was supplemented with a form of convenience sample that guided the selection of interviewees on the basis of whether they had the time to talk and whether they worked with different social dynamics in different organizational sectors.

Each case is represented by two types of data. The first is interviews with the project-leaders and the other is documents they linked to from their web sites or suggested as relevant before the interviews. The interviews were carried out between October 2011 and April 2012 in New York City, Boston and through Skype. They lasted between 45 minutes and 1 hour and the semi-structured interview-guides were inspired by the theoretical framework outlined above as well as the suggested documents. All interview guides alluded to digital choices and web-frames, but each of them had unique elements because of differences in the documents suggested by the interviewees. The transcribed interviews and the documents were coded and analyzed in NVivo and the specific data sources are listed in column three and four in figure 1. The code after the source will be used to indicate when they are referenced in the analysis.

| **Name of project-leader and organizational affiliation** | **Social dynamics of interest** | **Interview-data imported to NVivo** | **Document-data imported into NVivo**[1] |
|---|---|---|---|
| Ana Andjelic<br>Digital strategist and marketing consultant Droga 5 | Value creation and cultural tensions around brands | 35 minute interview (D1) | 2 years of blogposts by Ana Andjelic on the blog ´I [love] marketing´ (D2) |
| John Kelly<br>Co-founder and chief scientist at Morningside Analytics | Communities that share knowledge and focus attention on particular sources of information and opinion. | 1 hour interview (MA1) | 3 academic papers:<br>Pride of Place (MA2)<br>Mapping Irans Online Public (MA3)<br>Mapping the Arabic Blogosphere (MA4) |
| Alan Porter<br>Foresight analyst at Search Technology Inc. | Innovation paths around emerging technologies and trans-disciplinary reach of research fields. | 45 minute interview (STI1) | 3 academic papers:<br>Forecasting Innovation Pathways (STI2)<br>A Forward Diversity Index (STI3)<br>Assesing the Human and social dynamics program (STI4) |
| Chris Pallaris<br>Senior consultant at I-Intelligence | Open source intelligence and signals of changes that can support government policy and business strategy. | 1 hour interview (I1) | 1 academic paper:<br>OSINT – Knowledge, activity and organization (I2)<br>1 keynote presentation: |

[1] The documents can be obtained by contacting the author and their references are listed after the literature if they are not anonymized.

| | | | |
|---|---|---|---|
| | | | The Four Architectures of Competitive Intelligence (I3) |
| Vincent Lepinéy<br>Sociologist at MIT´s Mapping Controversies program. | The dynamics of socio-technical controversies | 1 hour interview (MC1) | NONE |
| Guilhem Fouetillou<br>CEO and co-founder at Linkfluence | Product-related conversations taking place in social web communities | 1 hour interview (L1) | NONE |
| [anonymized]<br>Founder and consultant at Information Service Bureau | Information-flows that can aid the quality of military intelligence. | 45 min interview (R1) | 1 keynote presentation<br>[xxxxx] (R2) |
| Robert Kirkpatrick<br>Director of the visualization branch ´Global Pulse´ at The United Nations | Early signals of crisis-related stress and other indications of developmental concern. | 1 hour interview (UN1) | 3 project white papers:<br>Twitter and perceptions of crisis related stress (UN2)<br>Using social media and conversations to add depth to unemployment statistics (UN3)<br>Streams of media issues: Monitoring world food security (UN4)<br>1 statement from the UN general secretary (UN5) |

Figure 1: Overview of case organizations

In the first part of the analysis (section 4), the coding of the empirical data was initially focused on identifying common challenges in constructing web-based visualizations. The codes resulting from this initial analysis broke the empirical data up into specific challenges that emerged as central across the cases. Because of the most different research design these challenges are taken as indications of influential constraints that need to be taken into account by organizations that engage in using web-based visualizations as an analytic device. The empirical data within each of these codes was then re-coded with the attempt of identifying the two most different approaches to meet these challenges across the cases. This led to the development of analytic ideal-types that indicate the flexibility that project leaders have in meeting the identified challenges. Being ideal-types these analytical constructs should not be seen as representing the position of the individual project leaders. They are rather providing insights into important trade-offs in the way analytical practices can be reconfigured through the use of web-based visualizations.

In the second part of the analysis (section 5), the coding way focused on the web-frames constructed in order to legitimize the digital choices taken to meet the identified challenges. Two themes emerged as central to the construction of such frames across the cases. One was epistemological assumptions about the proper way to benchmark data and the other was ontological assumptions about the world in which web-based visualizations can have a potential use as analytic devices. In combinations these two themes will be denoted as ´epistemic assumptions´. Within each of the identified themes the empirical data was once again coded for the most different epistemic assumptions across the cases. The detection of these differences allowed for indicating the way digital choices are influenced by organizational cultures and codified modes of thinking about knowledge creation. The web-frames constructed need to be acceptable in the context in which the visualizations are to be used.

The presentation of the results of the analysis below will provide conceptualizations of the common challenges, the ideal-type approaches to meeting them, the central themes in establishing web-frames around these choices and the different epistemic assumptions involved in the construction of such frames across the cases. Each concept will be illustrated by a few quotes from the empirical material but the style of presentation will focus on the analytical concepts that emerged from the analysis. The more detailed empirical story is, accordingly, traded for a focus on the theoretical concepts and the way they can indicate the socio-technical influences that will come to shape the future of web-based visualizations as an analytical device.

## 4. **COMMON CHALLENGES & DIGITAL CHOICES**

The first part of the analysis resulted in the identification of two common challenges that constrain and shape the way web-based visualizations are constructed and used as a device of social analysis across the cases. One is the necessity to *distribute part of the data-formatting* to third party actors and the other is the necessity to *balance the powers machine intelligence and human intuition* when it comes to automating the analysis of the collected data. Other challenges were mentioned during the interviews, such as the importance of aesthetically pleasing visualizations and the potential breach of data security. But the two challenges in focus were subjected to detailed analysis because they were reoccurring across the cases. In the presentation below they will briefly be described without reference to the empirical data and thereafter grounded in the empirical data through the presentation of the ideal-type approaches that are identified as the most opposite across the cases. This mode of presentation is chosen because the challenges are more visible in the explanations of the solutions than in explicit statements of the challenges themselves.

## 4.1. DISTRIBUTING DATA-FORMATS

To format data is here defined as the practice of segmenting it on the basis of a set of pre-defined specifications. A well-known example of such formatting is the way information on a web-site is segmented on the basis of distinctions between headlines, hyperlinks and anchor texts. Such specifications are the result of digital choices made in the early stages of the web and they have subsequently influenced the way search engines like Google analyze web-site content and decide upon its relevance (Battelle 2006). Clear and structured data-formats are necessary in order to get computers to process data and the digital choices involved in deciding upon them are decisive in shaping the final visualization. When looking at the empirical data it is clear that all the studied projects are involved in some sort of distribution of the practice of data-formatting to third party actors (see also Marres 2011). The digital choices taken in relation to this distribution are challenging because they provide access to data but at the same time involve a loss of control and transparency in the process of formatting it. Two ideal-type approaches to meeting this challenge were identified and they indicate a central trade-off in using web-based visualizations as an analytic device.

One ideal-type is conceptualized as the approach of *´channeling´*. It represents a suggestion to confine the distribution of data-formats to communication channels that are deemed valid and reliable in relation to sorting information about the specific topic of interest. This approach assumes the existence of specialized channels with unique competencies in formatting data from specific groups that communicate about specific issues through specific genres. ´Web of Science´ (WOS) is an example of such a channel and the visualization in figure 2 uses its data-formats as the basis for depicting the interdisciplinary reach of specific research practices. It is produced as an analytical device to help evaluate whether The U.S. National Science Foundation is succeeding in funding research that crosses disciplinary and organizational boundaries and it is explicitly argued to

"[…] depend on the WOS subject categories" (STI3). These categories are a specific way of segmenting information in scientific papers into pre-defined formats that can be distinguished from each other by a computer. These formats include author affiliations, citation scores, publication dates and journal categories (STI3; STI4). The visualization in figure 2 is built on these data-formats because WOS is a channel that has an explicit and institutionalized expertise in segmenting scientific texts. When a paper is classified as belonging to a specific category in WOS it is because a competent human with known competencies in the genre of scientific writing has placed it there. This does not mean that the formats are perfect but they are seen as sufficiently stable, well-defined and transparent to risk the loss of control involved in distributing decisions about them to a third party actor (STI3; STI4).

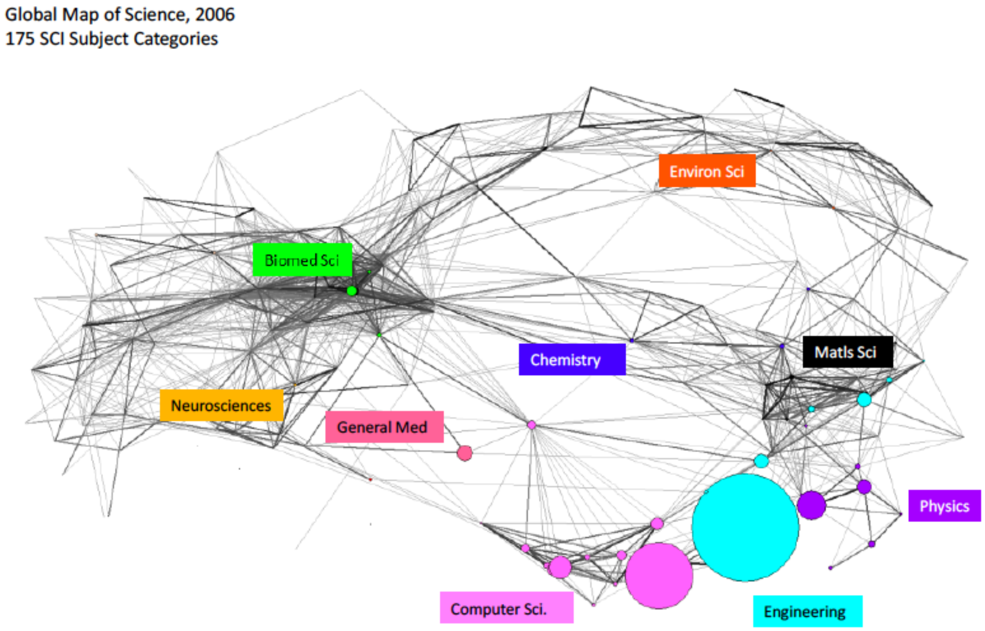

Figure 2: Visualization depicting the inter-disciplinary reach of scientific disciplines in order to evaluate whether The U.S. National Science Foundation is succeeding in funding research that crosses disciplinary boundaries

Characteristics of the approach of channeling are found across other projects as well. Data-formats in Thompson Reuter´s Derwent World Patent Index is, for instance, used as a basis for visualizing innovation pathways around emerging technologies in another project carried out by Search Technologies Inc. (STI2) and the formatting of press-releases in Dow Jones´ business tool, Factiva, is the basis for mapping the influence of different food security issues in a project carried out by Global Pulse (UN4). These projects contain elements of the ideal-type of channeling because they distribute the cognitive work of data-formatting to channels with an institutionalized expertise and a clear and transparent process for segmenting data. The rationale behind the approach of channeling is nicely summarized by the director of an Information Services Bureau who states that if one wants to know what, for instance, the medical profession think about a specific issue one must first look for "[…] whatever channel there is where medics discuss these things" (R1). This indicates that medics are the best sources of information about the medical profession and it is only if there does not exist specialized channels where medics communicate that one should turn to sources where the basis of the data formats is less transparent.

The other ideal-type is conceptualized as the approach of *´real-time tracking´*. It is an opposite approach to channeling because it takes advantage of the fact that new "[…] internet communications technologies are eliminating the channel-segregation" (MA2). Facebook and Twitter are examples of such technologies. They are not designed for communication between people with a pre-defined expertise that communicate in specialized genres. They are interfaces that "[…] function more as a media platform than as a publisher with editorial control" (D2; MA4). Such platforms provide a more diverse set of web-users with the opportunity to communicate and share information than, for instance, WOS. This makes it possible to, for instance, understand the spread of research without having to rely on data-formats built from within the discipline of science. A research idea would not necessarily have to come in the format of a paper made by an identifiable author and its uptake would not have to be judged on the basis of institutionally validated formats such as a citation. This makes data-flows fast and diverse but the trade-off is that the formats relied upon are determined by private companies. This means that they are neither transparent nor validated by any recognized institutions and they lack reliability because the interfaces on which they are left are constantly re-designed in order to increase traffic.

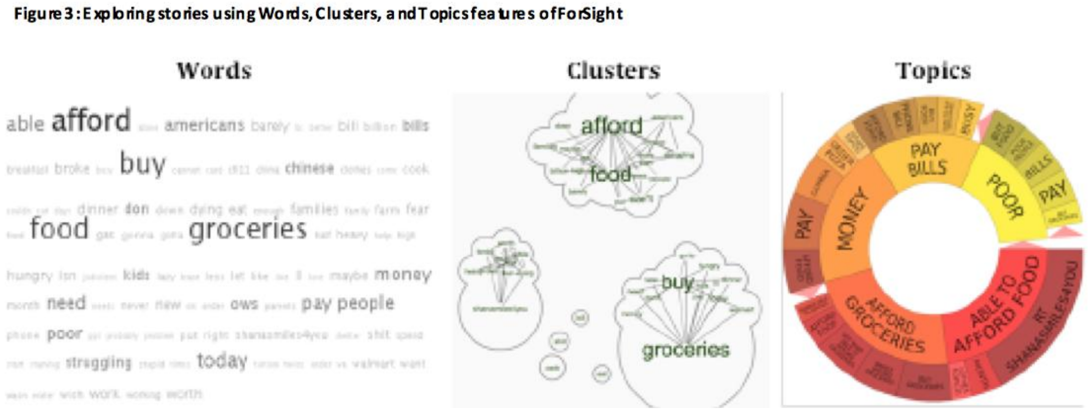

Figure 3: Visualization depicting meaning structures around the topic of food in order to detect negative emotions and early signals of crisis.

The visualization in figure 3 is guided by the approach of real time tracking because it trades the quality indicators of validity, reliability and transparency for fast and large data-flows. It is produced by Global Pulse and the data-format that provides its basis it the tweet. By harnessing semantic patterns in tweets it shows the meaning attached to the topic of ´food´ in Indonesia and USA. To the left we see the words most frequently used, in the middle we see the clusters of words used in combination with pre-defined key words and to the right we see a topic wheel that shows groups of related posts and the popularity of the topics they belong to.The visualization is produced in order to detect fast and early signals of vulnerable populations and crisis-related stress in the regions covered (UN2). But contrary to a traditional survey that would rely on formats such as pre-defined multiple choice questions it is not clear who the source of a tweet are, in which context it it left and what the motivations for leaving it were.

The tweet is, accordingly, not a valid and transparent format. It is chosen because the technical infrastructure around the practice of tweeting allows for ‘quick and dirty’ indications of crisis-related issues that are unmatched by, for instance, the format of a multiple-choice survey when it comes to the dynamics of data. By being limited to 140 characters a tweet ignites a culture of quick thoughts that is argued to “[…] reveal a great deal, particularly in the case of emergencies […]” (UN2). This is also true for the re-tweet which is repurposed in other projects. Compared with a structural indicator, such as the citation, the strength of the re-tweet is that it creates a dynamic culture of linking. Whereas the infrastructure around citations promotes a culture of persistent references, the re-tweet allows for a “[…] differentiation between the content space and the link space (L1)”. This enables people to show appreciation for specific aspects of a text rather than a whole paper. ´Real time tracking´ favors data-formats that make people constantly revise the traces they have left and make data-flows dynamic and responsive.

## 4.2 BALANCING MACHINE INTELLIGENCE & HUMAN INTUITION

Computers need data to be broken into recognizable segments such as a citation, an author affiliation or a tweet. But meeting the challenge of distributed data-formats gives rise to a second challenge that is common across the cases. This is the challenge of finding the proper level of trust in the machine intelligence needed in order to transform big data streams into comprehensible visualizations. The color and structure of visualizations are decisive for the way the attention of the reader is organized. The visualization in figure 2 is, for instance, colored on the basis of a category like ´biomedical sciences´ that is build though an automated analysis of overlaps between texts in pre-defined scientific practices (STI3; STI4). The coloring is, in that sense, distributed between the initial expert categorization discussed above and the algorithm running and inductive factor analysis. The way machine intelligence and human intuition is balanced is another digital choice where two ideal-type approaches have been identified to indicate another central trade-off in using web-based visualizations as an analytic device.

One ideal-type is conceptualized as the approach of *´following´*. It relies on a belief in algorithms as powerful tools that can recognize surprising associations and patterns without being distracted by cultural preconceptions. Algorithms are "blind" in their processing of data but proponents of following emphasize that one can be guided to innovative analytical concepts by following the blind. The visualization in figure 4 is, for instance, colored on the basis of algorithmic pattern recognition of the link histories of blogs (MA3; MA4). Whereas previous visualizations of the blogosphere have often been colored on the basis of pre-established distinctions between liberal and conservative bloggers it is here argued that intelligent algorithms can automatically "[…] locate these large political clusters as well as a number of other attentive clusters that […] prove to have their own thematic foci […]" (MA2). The resulting clusters does not map onto traditional categorizations about attention structures in the Arabic blogosphere, which is the social dynamic that the visualizations sets out to detect. The segmenting and coloring is therefore presented as a needed alternative to "[…] color the nodes on the basis of some pre-existing typology […]" (MA1). The ideal-type of following is, in that way, an attempt to bypass the drawbacks involved in relying too heavily on a priori human intuition (L1). It is an inductive approach that promises the readers of visualizations to see "[…] something that [they] have previously missed" (D2).

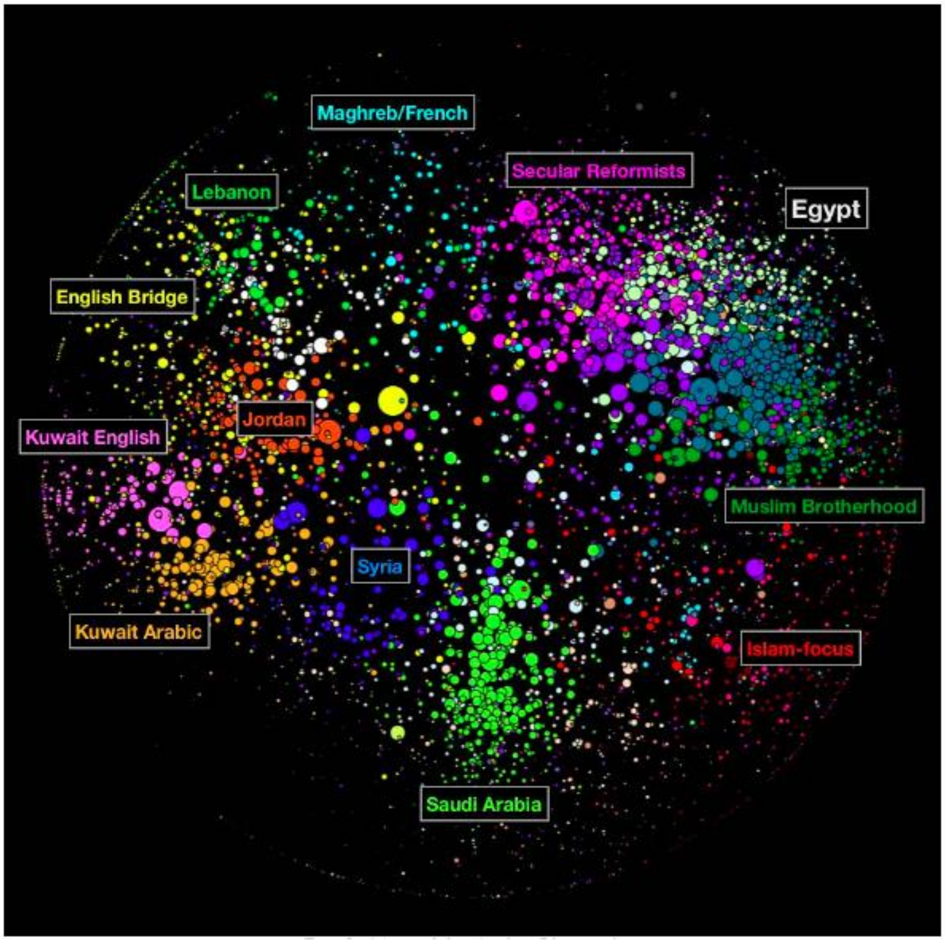

Figure 4: Visualization depicting ´attention clusters´ in the Arabic blogosphere in order to understand the influence of blogs on political discourse.

The opposite ideal-type is conceptualized as the approach of ´training´ because it emphasizes that "[…] it is imperative that the analyst "train the algorithm" […]" (UN2). It is an alternative to following in the sense that the algorithm in guided by a priori categories in order to make sure that it returns meaningful and useful visualizations. This is a way to prioritize the unique competencies of human intuition in interpreting semantics and social dynamics and it is argued that it is "[…] really difficult to have good results with purely automated approaches" (L1; R2; MA1). The tag-clouds in figure 3 above are built in accordance with the ideal-type of training in the sense that the underlying algorithms are constructed to detect emotions around pre-defined crisis-categories such as ´food´. Tweets that fit the intuition of the analysts about what belongs to this category are used to optimize the algorithms and ensure that the visualization is "[…] aligned with project objectives (UN2). The strategy behind the approach of training is, accordingly, to take advantage of the fact that computers have a processing capacity that dwarves that of humans but acknowledge that computers lack semantic skills (MA1). Ultimately it is a way to emphasize that "[…] at the beginning you always have a human decision (L1)". This decision is based on human intuition that points the "[…] processing capacity at particular problems" (I1).

## 5. WEB-FRAMES & THE LEGITIMIZATION OF DIGITAL CHOICES

The challenges and approaches outlined in the section above indicate some technical constraints that the construction of web-based visualizations must be practiced within. One example is that the relevant data-flows are often happening on technical infrastructures owned by third parties and this limits the control and transparency of data-formats. Another is that the flow of data is so fast and large in scope that the analyst needs to distribute important parts of the analysis to machines that have certain limitations in the way they process, for instance, semantic data. These technical constrains influence the digital choices taken in the construction of visualizations but they are not the only influences. Digital choices are also shaped by epistemic assumptions held by the project-leader and the organizational environment within which the visualizations are to be used. The fact that a visualization is technically feasible does not mean it is perceived as useful.

This section will focus on the web-frames that are constructed in order to make web-based visualizations sensible and legitimate and it identifies two themes that are central to the construction of web-frames across the cases. One is the need to establish a point of reference against which to evaluate the validity of digital data and make it ´hard´ enough to be a legitimate basis for decision making. This theme illustrates how epistemological assumptions about the proper way to *benchmark data* can shape digital choices. The other theme is the need to build visualizations on the basis of underlying assumptions about the stability of the world in which they are to serve as analytic devices. Such ontological assumptions concern the *conditions of decision-making.* They guide the problems that visualizations are intended to solve and thereby also the digital choices made in their production. Within each of these themes the data has, once again, provided the foundation to construct two ideal-type positions that shape digital choices in different ways. As above, it is in the presentation of these ideal-types that the identified themes get empirical grounding.

## 5.1 BENCHMARKING DATA

One ideal-type position in relation the theme of data-benchmarks is conceptualized as *´external correspondence´*. This position builds on the assumption that data can only be validated through the existence of benchmarks that are external to the tools through which it is produced. Since the primary tools used to build web-based visualizations are software crawlers that process digital traces it is not surprising that the position of external correspondence is characterized by looking for some sort of offline baseline with which to evaluate the validity of the digital data produced. This requirement surfaces in two forms across the cases.

One form is the reliance on *expert validation* that we have already encountered in the approach of channeling. It relies on the epistemological assumption is that data is a legitimate basis for decision-making when it is transparent enough for a competent expert to trace it back to its source and evaluate its validity. It is only when the competencies of experts are clearly identified and when data is transparent that it is possible to provide a valid translation of the external world into bits of data to be processed by a computer (STI2). This criteria of data-quality is even quantified in the statement that the anonymity of a source makes its ”[…] information value go down with 50 percent […]” and lack of knowledge about the source makes it go “[…] down with another 25 percent” (R2). If one is meeting the request for external correspondence through the approach of expert validation one is, accordingly, focused on identifying honest brokers of verified information to legitimize the data.

The other form in which the assumption of external correspondence surfaces across the cases is substituting the focus for honest brokers for a focus on honest signals. The idea is that such signals can be obtained through *minimal human interference* in the process of translating raw data into web-based visualizations. The underlying epistemological assumption is that human involvement in the process of data-selection is a source of bias rather than a source of validity. Digital traces are seen as useful data on which to build strategies precisely because the people that leave them are not obstructed by a researcher. They are argued to represent “[…] spontaneous conversations” (L1) and they are argued to produce situations where people are “[…] broadcasting how they feel what they do and what they think” (D1). This is explicitly contrasted to methods like focus groups that are seen as “[…] artificial environments […]” (D1). Web-based visualizations are conceived as legitimate analytical tools because they have the potential to "[…] reflect our inner human nature" (I1) and because they are not "[…] based on inferences" (MA3). They correspond to the true offline world because there is no researcher bias involved in the translation from the way people think and behave to the digital traces signifying these thoughts and behaviors.

The opposite ideal-type position in relation the theme of data-benchmarks is conceptualized as *´ pragmatic coherence´*. It is an alternative to evaluate data on the basis of whether or not its correspondence to some external phenomenon is ensured by expert validation or minimal human interference. The underlying assumption is that digital traces are neither transparent nor honest. They are always biased and messy. When, for instance, Global Pulse use tweets as signals of crisis-related stress in figure 2 it is explicitly acknowledged Twitter is a platform that has "[…] a specific culture and demographic [that] change over time and varies by topic, location, and other factors" (UN2). Tweets are seen as cultural products rather than honest signals but they are still argued to be legitimate and actionable data. Their legitimacy is, however, dependent on a solid knowledge of the specific culture around their production (UN1). Pragmatic coherence simply builds on the assumption that it is possible to construct pragmatic benchmarks that are coherent with this culture and its potential biases.

This open for the construction of benchmarks that are internal to the tools that produce the data and the approach of the Global Pulse project is simply that "[…] the most straightforward analysis [is] based on daily anomaly detection" (UN2). Sensitivity towards anomalies is presented as the key to utilize of the intelligence potential of digital traces without slowing their use down by adhering to traditional quality indicators such as correspondence some external benchmark. The ideal-type of pragmatic coherence is resurfacing in other projects that explicitly accept that "[…] people don't act the same way online [as] in their real life and [they] won't say exactly what they think […]" (L1). In short it is accepted that "[…] you don't have access to their intimate representations and thoughts" (L1). This is, however, not seen as a problem by proponents of pragmatic coherence. Bias, lack of transparency and the influence of platforms on people's behavior is turned into a source of insight once the analyst has enough knowledge to build coherent internal benchmarks with which to detect anomalies in the flow of data.

´Pragmatic coherence´ is accordingly an ideal-type position that represents an attempt to re-configure the epistemological foundations of social analysis by questioning inherited epistemic assumptions. Judged from the perspective of external correspondence such a re-configuration would have several deficiencies. Proponents of expert validation would point to the fact that Twitter is a nontransparent source that only makes a subset of its data available for analysis through their Application Program Interface (APIs). The criteria for selecting this subset are furthermore in-transparent and it is even an open question whether information is censored before it becomes available though the API (R1). The data is fianlly anonymous and bound to a pre-defined format of 140 characters which leaves very little space to provide a true depth of feeling (I1; MC1). Proponents of minimal interference would be especially uneasy with the way proponents of ´pragmatic coherence´ seems to take non-honesty of data as an uncontroversial fact.

## 5.2 GROUNDING DECISON-MAKING

Approaches to data-benchmarks are tightly coupled to assumptions about the ontological status of the world because it determines the conditions under which decision-making takes place. The need to build visualizations on ontological assumptions about the world is the other central theme in the construction of web-frames an ideal-type position that is strong across the cases is conceptualized as a belief in *fluid social realities.* The central assumption underneath this position is that new communication technologies are "[…] changing the nature of information [in a way that] reflects a larger, structural remaking of society whose end state we cannot predict" (I2). The central argument is that the world has become "[…] interactive, networked, info-rich [and] collaborative" (D2) in a way that confronts analysts and strategists with "[…] situations of uncertainty [and] ambiguity" (I2). Proponents of ´fluid social realities´ argue that this ontological change must be reflected in the analytical tools we use to understand it.

The Secretary-General of the UN has recently exemplified this position by stating that the world is increasingly "[…] volatile and interconnected [because] the impacts of [a] crisis [is] flowing across borders at unprecedented velocity" (UN5). A consequence of this is that tools like surveys and census data are too slow to detect signals of emerging crises in due time to react upon them. An example of such a signal is a mother that takes her kid out of school. Within the UN this is considered to be an ´early signal´ of economic problems and the argument for revisiting analytic methods is that this mother will communicate about her choice through a traceable media-device a long time before a traditional survey can capture it (UN1). It is such experiences that makes the Secretary-General conclude that "[…] traditional 20th-century tools for tracking […] development simply cannot keep up […]" (UN5).

All projects echo the idea that surveys and census-data do "[…] no longer have a monopoly on the knowledge" (I2) and another central assumption underlining the position of fluid social realities is that relevant data must be found across a range of sources and actors (D2). The new world of information means that a diverse set of actors have the competencies to browse through data and reveal deficiencies in the way organizations have treated the data. When, for instance, a crime is committed it will often be possible to backtrack the behavior of the criminal through their digital traces and find signals that could have led to a capture before the crime. Proponents of fluid social realities emphasize that this creates an increasing demand for "[…] more intelligence, more quickly, and more often" (I2). This has the consequence that "[…] short-term situational assessments will likely be given preference over long-term strategic projections" (I2). Digital choices made in the construction of web-based visualizations must therefore differ from the methodological choices made in a situation where confidential and validated data was used to make long-term projections.

The position of ´fluid social realities´ is heavily dominating the cases and the conceptualization of an alternative ontological position is to a large extent based on negative definitions. An alternative position can be conceptualized as a belief in *strategic continuity* and it is widely talked about as the ontology of the past by proponents of fluid social realities. They see the persistence of this position as an obstacle to make web-based visualizations legitimate analytic devices because it clings to a belief in stable classifications and competencies that favors other methods. It is claimed that proponents of strategic continuity often agree that the Internet has changed flows of information but they fail to see how this entails a new ontology that requires radically new analytical devices and modes of organizing. The position of strategic continuity is seen as reflecting ´organizational pathologies´ that make it hard for new analytical methods to travel. One example of such a pathology is the automatic belief in stability that, for instance, characterized most policy reactions to the Arab spring (I1). Another is the tendency rely on ´silos of expertise´ in a world where data-competencies are distrubuted. A third is the adherence to ´bureaucratic thinking and hierarchies´ that treats information as secretive (I1). Such pathologies are argued to entail an unproductive belief in established cultures of evidence and a failure to, for instance, recognize "[…] that Youtube is a valid source of information" (I1).

There are, however, examples of statements that seem to support the position of strategic continuity in some of the projects. We have already encountered assumptions about stability in channels and contexts of expert knowledge and we see signs of it in the statement that "[…] analysis is explaining why something has happened [and] predicting what might happen in the future" (R1). The belief that prediction is a thing of the past is accordingly not unchallenged and such a belief is necessarily accompanied by assumptions about a certain level of stability in the world that is the object of prediction. The position of strategic continuity reflects a reluctance to re-configure analytical workflows and methods in response to the rise of technologies connected to the Internet. As put by one of the interviewees; “The search strategies remain the same despite the information format […]” (R1). The assumption is that the methodological foundations of analysis will remain the same even though we have entered a digital world. But all of the cases contain elements that challenge the approach of ´strategic continuity´ and it is evident that it represents a mode of thinking that is widely seen as obstructive to the uptake of web-based visualizations as analytic devices.

## 6. BETWEEN TECHNICAL CONSTRAINTS & EPISTEMIC ASSUMPTIONS

”Organizations are accounts of the change that is happening around them” (D2) argues one of the interviewees. If this is true it could be added that such accounts are heavily shaped by the analytic devices through which they make sense of this change. Web-based visualizations are increasingly used as such a device in a diverse set of organizations and their final shape is ultimately tied to the digital choices taken in the process of their construction. The analysis above have illustrated how such choices must be made within the context of some general challenges that all visualization projects share but also how specific digital choices are flexible and influenced by epistemic assumptions in the context in which the visualization is to be used. In combination these analytical results indicate that digital choices are influenced by two quite different mechanisms. One is the technical constraints involved in making big data sets apt for computerized analysis and the other is the need to ground digital choices in web-frames are coherent with epistemic assumptions in the organizations (and societies) within which they are meant to travel. As the director of Global Pulse puts it when taking about the potential for methodological innovation in within the UN: “[…] It is not just about getting the data; it is also […] about the organizational capacity to facture a snapshot of these types of information in the context of their on-going policy development planning” (UN1). The rest of the paper will revisit the main results of the analysis by reflecting on the insights they can provide in relation to analyzing visualizations that are put forward as devices of policy-planning.

The conceptualization of common challenges in section 4 can be seen as an indication of the most important technical constraints on visualizations introduced as devices for policy planning. The first challenge was the need to distribute the practice of data-formatting to external partners. The digital choices made in meeting this challenge were shown to be constrained by the need to accept the current power-structures in digital data ownership, the need to cope with rapid changes in the technical infrastructures on which data is left, the limitations of using API´s to harness data and so on. Organizations that use web-based visualizations as an aid for policy planning must accordingly find ways to act in a situation where they do not have the technical means to collect, analyze and distribute the relevant data themselves. The second challenge was the need to balance machine intelligence and human intuition in the processing of data. The digital choices made in meeting this challenge were primarily shown to be constrained by limitations in the way computers understand semantics. The constraints accompanying these common challenges will shape the digital choices made in the construction of web-based visualizations and therefore also the way they can be used as devices for improving, for instance, policy-planning.

It was, however, also shown that there is flexibility in making these digital choices and this indicates different potential ways of using visualizations as a policy device. The scope of variation was illustrated through the conceptualization of ideal-type approaches to deal with the identified challenges. A project that prioritizes expert validation and transparency in the distribution of data-formatting will, for instance, result in quite different policy-guidance than a project that accepts low transparency in the favor of real-time tracking. Looking at the challenge of automatization there are, similarly, differences between projects that ´follow´ algorithms and projects that ´train´ them. The first approach promises innovative categorizations whereas the latter ensures that the visualizations returned are immediately recognizable within the organization in which they are to be used. The digital choices involved in these approaches are identified as central to the way web-based visualization will guide the attention of organizations using them. The trade-offs involved in taking different positions on these choices will be central to the discussions about the legitimacy of using visualizations as a device for policy-planning as well.

Section 5 focused on the topic of legitimacy by showing how digital choices are also shaped by the web-frames constructed to align them with epistemic assumptions in the organizations and societies in which they are to be used. This aspect is crucial in relation to their potential success as policy devices and two themes were conceptualized as central in the construction of web-frames. One is the need to benchmark data and the other is the need to provide an ontological description of the status of the world in which web-based visualization are to function as strategic devices.

The analysis identified two different ideal-type positions on the theme of benchmarking. One was to assume the need for data-benchmarks that are external to the used software tools in order to verify their results. This quest for external correspondence was argued to build on a distinction between the ´virtual´ and the ´real´ and it illustrates how inherited quality-indicators can potentially shape the choices taken in the construction of visualizations. Other projects confronted such quality indicators by arguing for the legitimacy in constructing benchmarks that are internal to the tools used to build the visualizations. Rather than looking to offline data as the ultimate yardstick of good data they suggested to utilize the intelligence potential of the web by accepting that all benchmarks are built on shaky grounds (Rogers 2009). Finding coherent patterns that allows for detecting anomalies in web-data was argued to be the only way to harness the intelligence potential in digital traces without slowing the data-flow by important traditional quality-criteria. The epistemological discussions arising from the difference between these positions is another decisive dynamic that will shape the way web-based visualization can be used as devices of policy-planning.

Ontological descriptions of the world were shown to be necessary in order to establish a common idea of the conditions under which web-based visualizations can function as analytic devices. These descriptions showed a more unified set of assumptions across the cases. Most argued that there is a need for new analytical tools if we are to understand an increasingly fluid social reality. The broad adherence to the assumptions of fluid realities indicates that web-based visualization projects are to a large extent grounded in a tradition of incremental and emergent ´strategizing´ (Mintzberg 1990) that provides an alternative to rational and deductive approaches to strategy and analysis (Porter and Kramer 2006). This view on strategy and decision-making seems to indicate the direction that web-based visualizations can potentially move the practice of policy-planning.

In summary, it can be argued that the steps towards making web-based visualizations a useful policy-device must be approached as a socio-technical process. This process is first of all shaped by challenges that set some general technical constraints on the way such visualizations can be shaped and thereby how they can potentially guide attention to relevant social dynamics. But this does not mean that re-organization of analytic workflows can be seen as an output of technical affordances. The fact that a visualization is technically feasible does not mean it will be used. Such a use requires the establishment of a legitimate web-frame around it. Existing epistemic assumptions are in that way shaping the digital choices taken in constructing web-based visualizations as well. These outcomes of the analysis should be seen as a small step towards more solid conceptualization of the re-configuration of analytical methods that the use web-based visualizations have begun. The identified challenges and themes indicate the overall frame within which this reconfiguration takes place and the identified ideal-types indicate important lines of disagreement as to how web-based visualizations can be constructed within these challenges. Such conceptualizations are important guidelines when studying the way web-based visualizations are currently re-configuring analytical workflows within the field of policy-planning.

## References:

Anderson, Chris. 2008. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete", *Wired Magazine,* 16.07.

Battelle, John. 2006. *The Search - How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture*, New York: Portfolio

Brin, Sergey & Larry Page. 1998."The anatomy of a large-scale hypertextual Web search engine" *Computer Networks and ISDN Systems*, 30 (1-7): 107-117.

Dutton, William H. 2007. "Reconfiguring Access to Information and Expertise in the Social Sciences: The Social Shaping and Implications of Cyberinfrastructure". http://www.ncess.ac.uk/events/conference/2007/papers/paper152.pdf. (July 15 2009)

Flor, Grace d. & Eric Meyer. 2008. "Talking 'bout a revolution: Framing e-Research as a computerization movement". http://ora.ouls.ox.ac.uk/objects/uuid%3A031cf6ec-c2d5-4864-8f20-271588b221ab. (July 16 2009)

Flyvbjerg, Bent. 2004. "Five misunderstandings about case-study research" *Sosiologisk Tidsskrift*, 12(2): 117-42

Lazer, David et. al. 2009. "Computational Social Science" *Science Magazine*, 323: 721-723.

McKinsey Global Institute. 2011. *Big data: The next frontier for innovation, competition, and productivity*. http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation. ( August 1 2012)

Madsen, Anders K. 2012."Web-visions as controversy-lenses" *Interdisciplinary Science Review*, 37(1): 51-68

Manovich, Lev. 2008. "Software takes command". http://lab.softwarestudies.com/2008/11/softbook.html. (January 20 2012).

Marres, Noortje. 2011. "Re-distributing methods: digital social research as participatory research, Sociological review", Goldsmiths research online

Marres, Noortje. forthcoming. “The Experiment in Living.” In *Inventive Methods: The Happening of the Social*, ed. C. Lury and N. Wakeford. London: Routledge.

Mintzberg, Henry. 1990. “The Design School: Reconsidering the Basic Premises of Strategic Management”, *Strategic Management Journal*, 11(3): 171–95.

Porter, Michael E. and Mark R. Kramer. 2006. ”Strategy and Society”, *Harvard Business Review*, 84(12): 78-92.

Rogers, Richard. 2009. *The end of the virtual: Digital methods*. Vossiuspers UvA

World Economic Forum (2012), *Big Data, Big Impact: New Possibilities for International Development*. http://www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_2012.pdf. (August 1 2012)

## Documents used as empirical data to supplement the interviews:

Andjelic, Ana. 2010-2012. “I [love] Marketing”. Blog. http://anaandjelic.typepad.com/ (February 27 2012)

Global Pulse. 2011. “Streams of Media Issues - Monitoring World Food Security”. http://www.unglobalpulse.org/projects/news-awareness-and-emergent-information-monitoring-system-food-security (January 11 2012)

Global Pulse. 2011. “Using social media and online conversations to add depth to unemployment statistics“.http://www.unglobalpulse.org/projects/can-social-media-mining-add-depth-unemployment-statistics (January 11 2012)

Global Pulse. 2011. “Twitter and perceptions of crisis related stress”. http://www.unglobalpulse.org/projects/twitter-and-perceptions-crisis-related-stress (January 11 2012)

Kelly, John et.al. 2009. “Mapping the Arabic Blogosphere: Politics, Culture, and Dissent”. Berkman Center Research Publication. http://cyber.law.harvard.edu/publications/2009/Mapping_the_Arabic_Blogosphere (November 13 2011).

Kelly, John and Bruce Etling. 2008. “Mapping Iran´s Online Public: Politics and Culture in the Persian Blogosphere, Berkman Center Research Publication. http://cyber.law.harvard.edu/publications/2008/Mapping_Irans_Online_Public/ (November 13 2011).

Kelly, John. 2008. “Pride of place: Mainstream Media and the Networked Public Sphere, Berkman Center Research Publication. http://cyber.law.harvard.edu/sites/cyber.law.harvard.edu/files/Pride%20of%20Place_MR.pdf (November 13 2011).

Ki-Moon, Ban. 2011. “Secretary-General's remarks at General Assembly Briefing on the Global Pulse Initiative. http://www.un.org/sg/statements/?nid=5668 (January 11 2012)

Pallaris, Chris. 2009. “OSINT as Knowledge, Activity and Organization- Trends, Challenges and Recommendations”. http://www.eurosint.eu/system/files/docs/osint-knowledge-activity-organization.pdf (February 27 2012)

Pallaris, Chris. 2011. “The Four Architectures of Competitive Intelligence, presentation at 360 Degree Indian CI Conference (February 27 2012)

Porter, Alan and Stephen Carley. 2011. “A Forward Diversity Index”. *Scientometrics,* 90(2): 407-427

Porter, Alan et al. Forthcoming. ”Forecasting Innovation Pathways: The Case of Nano-enhanced Solar Cells”, Technological Forecasting & Social Change

Porter, Alan and John Garner. 2011. “Assessing the Human and Social Dynamics Program□Exceptional Cross-disciplinarity. Paper presented at Atlanta Conference on Science and Innovation Policy.