This is a repository copy of *Towards an understanding of data work in context: Emerging issues of economy, governance, and ethics* .

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/95882/

Version: Accepted Version

Article:
Foster, J.J. (2016) Towards an understanding of data work in context: Emerging issues of economy, governance, and ethics. Library Hi Tech, 34 (2). pp. 182-196. ISSN 0737-8831

https://doi.org/10.1108/LHT-12-2015-0121

Reuse
Unless indicated otherwise, fulltext items are protected by copyright with all rights reserved. The copyright exception in section 29 of the Copyright, Designs and Patents Act 1988 allows the making of a single copy solely for the purpose of non-commercial research or private study within the limits of fair dealing. The publisher or other rights-holder may allow further reproduction and re-use of this version - refer to the White Rose Research Online record for this item. Where records identify the publisher as the copyright holder, users can verify any specific terms of use on the publisher's website.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Towards an Understanding of Data Work in Context: Emerging Challenges for the Data Professional

Structured Abstract:

Purpose. It is a commonplace that innovation in the digital economy is now driven by data. Business organizations, media companies, and government for example all create economic and societal value from the digital traces left by the user population. At the same time the data captured also contains information that personally identifies consumers, citizens and patients as individuals. The purpose of this paper is to place this new form of data work in the context of previous approaches to information work; to identify the differences between information and data work and the resulting challenges for data professionals.

Design/methodology/approach. Informed by a review of previous approaches to information work, the article argues that the shift in value from information to data as an economic asset and a societal good entails a new form of human-oriented data work. One that is more sensitive to the contextual conditions and consequences of the capture, processing and use of data than has been the case hitherto. The implications of this for a shift in emphasis from the data scientist to the data professional is addressed, as are emerging challenges of governance and education.

Findings. The main consequence for data professionals is to ensure that processes are in place not only to enable the creation of valued products and services from data, but also to mitigate the risks related to their development. The paper argues that ensuring this involves taking a contextual view that locates data processing within the user, governance, legal, and ethical conditions related to data work. The consequences for the governance of data, and the education of data professionals are addressed.

Originality/value. The value of the paper rests in its development of an analytical and methodologically driven framework, that places new forms of data work in the context of their conditions and consequences. The framework builds on prior approaches to information work, current approaches to data work, and addresses the governance, and educational challenges arising from organisations' emphasis on data-driven innovation in a digital economy.

Introduction

In order to survive and flourish, all organizations require an adequate understanding of the environments within which they operate (Choo, 2001). This has typically involved members scanning the organization's external and internal environments, and engaging in the planned and deliberate seeking of information, based on management information needs. In an era of big data, organizations are not only engaging in environmental scanning but are also leveraging the digital traces left by customers, clients, consumers, citizens and others as they interact with organizations via the web, tablets, and smartphones. Instead of information being sought and pulled from the environment, data is being pushed at organizations at a scale that was

unimaginable even a few years ago. These data include the involuntary collection of browsing and search data, location-based data, sensor data, and other personal identifying information (PII) that are automatically captured by the platforms and services that we use; along with the processing of other voluntary textual, aural, and visual information that we explicitly contribute via blogs, opinion sites, and social media etc. Big data has a number of attributes including its volume, velocity and variety (Laney, 2001). Volume: online channels increase the depth/breadth of data that can be collected on a given transaction or point of interaction; big data are also big in terms of enabling the capacity to look for patterns at new levels of scale. Velocity: increases in point-to-interaction speed are increasing the quantity of temporal data available e.g. real-time analytics. Variety: The variety of data sources from which data is captured includes search systems, webpages, clickstreams, social media logs, customer relationship management and other systems. The capacity to aggregate information about the preferences, actions, and behaviours of individual system users, to make connections across these different streams of data – and thereby add value – is a complex task involving questions of accuracy, standards, and verifiability. In short the capacity to turn not only information but also data into an economic asset and societal good is fast becoming part of all organizations' core competence; whether it be a business, a social media company, a government department, hospital, educational establishment, or scientific research institute. The organizational motivations for innovating with data are clear, e.g. personalization, community-building, product development and service improvements. However there is also a requirement, and a duty, for organizations to mitigate the risks that data-driven innovation poses to users and to organizations. The implications of data work for data professionals has received insufficient coverage in the information science literature. The structure of the paper is as follows. In a first section, some of the prior approaches that have been used to define information work and to understand its nature and scope are reviewed. In a second section, some of the developing approaches to organizing data work are then reviewed. In a concluding section, observations are made about the implications of data work for data professionals.

Approaches to Information Work

A clear connection between information work and the economy has undoubtedly always existed, since previous studies and definitions of information work have for the most part been developed within an economic context, e.g. evaluating the contribution that information work makes to economic productivity, or costing an organization's information function. Within this economic context, a number of approaches can be distinguished. First, sociological and occupational approaches that rely on a structural understanding of the number of workers involved in the production and analysis of information (Bell, 1973; Reich, 1991; Webster, 2006). Second, approaches that rely on an understanding of the organizational practices and activities surrounding information production as a primary good, or as a secondary good in support of the production of other primary goods (Porat, 1977; Hardt and Negri, 2000; Benkler, 2006; Foster, 2013). Finally, approaches that rely on understanding information as an economic asset (Hawley Committee, n.d.; Horne, 1995; Oppenheim, Stenson and Wilson, 2003a, 2003b, 2004a; Wilson and Stenson, 2008). In these last studies the common theme is to identify what information assets the organization holds, measure the costs involved in their acquisition or production, and establish their benefits for the organization. A review of each of these approaches is provided, before consideration is given to how an understanding of new forms of data work entails some continuation of, but also some change, in our understandings of information work.

Occupational Approaches to Information Work

Webster (2006) identifies five criteria developed to support arguments in favor of the emergence of an information society: technological, economic, occupational, spatial, and cultural criteria. He also adds a further criterion in the form of expert knowledge. For reasons of relevance the occupational criterion is focused on here. According to this criterion, an information society can be said to have emerged when a quantitative shift towards information work has occurred in the occupational structure of work. In other words, when a quantitative shift away from manual work towards jobs where the manipulation of information is the key task. For example, computing, accounting and finance jobs as well as those involving the production of media content etc. The critique of the occupational definition, as Webster points out, is that it is based on a

distinction of degree, not of kind. In other words it is not based on a distinction between the kind of work undertaken, but the degree of manual vs. informational work undertaken as part of the same job. For there are many occupations that combine both aspects within the same job e.g. railway signalman, lighthouse keeper. The question then becomes, are these jobs to be counted as consisting of manual work or information work? Nevertheless the occupational criterion has achieved wide currency, and was taken up by Robert Reich, US Secretary of Labor during the 1990s Clinton Administration. Schematizing the nature of labor in 1990s America Reich (1991) identifies three main categories of work: routine production services, in-person services and symbolic-analytic services. Routine production services consist of the repetitive operational tasks performed in high volume enterprises in order to produce the final goods required. These tasks are typical of foremen, line managers etc. but also include repetitive supervision. The category of in-person services also involves repetitive tasks, but these are distinguishable from routine production services due to the direct contact with the people who benefit from the services that the jobs involve. Therefore a key requirement differentiating in-person service personnel from routine production workers is that they are required to have "a pleasant demeanor" (Reich, quoted in Webster, 2006: 206). Retail sale workers, hotel workers, cashiers, home health care aides, hairdressers, flight attendants etc. can be counted as examples of jobs in this category. The third category of work consists of symbolic-analytic services. These jobs have a different goal being concerned neither with the production of material things nor with human contact, but with the "manipulation of symbols – data, words, oral and visual representations" (Reich, quoted in Webster, 2006: 207). The occupational criterion is also one implicitly used by proponents of immaterial labor, "the passage toward an informational economy necessarily involves a change in the quality and nature of labor…today information and communication have come to play a foundational role in production processes […] The service sectors of the economy [also] present a richer model of productive communication. Most services indeed are based on the continual exchange of information and knowledge. Since the production of services results in no material and durable good, we define the labor involved in this production as *immaterial labor*, such as a service, a cultural product, knowledge or communication" (Hardt and Negri, 2013: 289-290). While Hardt and Negri's characterization of immaterial labour shares similarities with Webster (2006) and Reich (1991), their definition is notable

for incorporating an affective element, arguing that three types of immaterial labor exist: "informationalized industrial labor, symbolic and analytic labor, and production and manipulation of affect [which] requires (virtual or actual) human contact, labor in the bodily mode" (Hardt and Negri, 2000: 293). Over the past two decades, the emergence and maturity of the Internet, tablets and smartphones, has extended immaterial labor from organizations to society, such that we can talk of a 'social factory' where all members become involved in the production of both informational and non-informational goods. In this respect the relations between organizations on the one hand, consumers and the public on the other has become a key strategic arena for capitalism. Foster (2013) identifies a number of ways in which consumers and members of the public involve themselves in valorizing the cultural content of goods e.g. via peer or individual production of information, via the consumption of 'free' content, or via textual, aural or visual 'utterances'.

Organizational Approaches to Information Work

Set against the backdrop of an emerging post-industrial society – and hence an interest in the economic importance of knowledge communication and information – Porat (1977) asks the following question: "*What share of our national wealth originates with the production, processing and distribution of information goods and services*"? *Or, what is the extent of the information activity,* (*as opposed to agriculture, services or industry*)*, as a portion of the total U.S. economic activity*" (Porat, 1977: 1-2). In developing an answer to this question, Porat makes an initial distinction between two economic domains; one concerned with generating wealth by transforming matter and energy from one form into another, the other concerned with generating wealth from transforming information from one pattern into another. However, much like the occupational definition, Porat does not see the industrial and informational domains as mutually exclusive. Indeed a useful contemporary illustration of their interrelationship is the emergence of 3D manufacturing where instructional code and materials are combined as elements within the same production process. Having attempted to distinguish the economic domain of informational value, Porat then proceeds to identify 'information activity' rather than information *per se* as the key unit of analysis. It is not information as an object or thing, "data that have been organized and communicated", that is economically productive but rather a more complex 'information activity': "Information is not a homogeneous good or service

such as milk or iron ore. It is a collection or a bundle of many heterogeneous goods and services that together comprise an activity in the U.S. economy. For example, the informational requirements of organizing a firm include such diverse activities as research and development, managerial decision-making, writing letters, filing invoices, data processing, telephone communication, and producing a host of memos, forms, reports, and control mechanisms" (Porat, 1977: 2). In other words the "information activity includes all the resources consumed in producing, processing and distributing information goods and services". These resources comprise information capital (or what can be termed fixed capital) and information labor (or what can be termed variable capital). 'Information capital' consists of "resources…used to deliver the informational requirements of one firm: typewriters, calculators, copiers, terminals, computers, telephones and switchboards. And depending on the size of the firm, there could be a massive array of high technology information goods such as microwave antennae, satellite dishes, and facsimile machines. On the labor side, the firm has to employ the services of many different types of 'information workers', who together satisfy the firm's informational requirements. We find the research scientist, engineer, designer, draftsman, manager, secretary, clerk, accountant, lawyer, advertising manager, communications officer, personnel director – all essentially paid to create knowledge, communicate ideas, process information – in one way or another transform symbols from one form to another" (Porat, 1977: 2-3). In summary, according to Porat, and in keeping with the service ethos of a post-industrial society, the economic value of information rests less in information content, and more in how informational activities contribute to the production of information and other types of goods and services.

A more recent approach to information production is presented in Benkler (2006), who proposes that a networked information economy has emerged in the area of information goods and content, e.g. text, news, films and music in digital form, software code etc. In a networked information economy, these information goods can be produced in a number of different ways. Historically the only strategy used to produce information goods has been proprietary and market-based e.g. the author writes the book, enters into a contract with a publisher, the work is then copyrighted and sold via the market. In this way the authoring of the work is incentivized via a monetary payment. However a digital, networked, information production strategy is

able to take advantage of the intangible, non-rival, nature of information goods. First, one person's consumption of a book, an item of news, a film does not necessarily diminish the opportunity of others to consume those same information goods. Secondly, once a digital version of the goods has been created, e.g. an e-book, very few additional societal resources need be consumed to produce additional copies to satisfy demand. In this way a proprietary, non-market, information production strategy has emerged in which organizations' subsidized production and distribution of free versions of the goods first engages the attention of the user; while the more complete version or service monetizes this initial attention at a subsequent point in time. Democratization of the means of information production, e.g. Mac, PC, coupled with the Internet as a distribution platform, has also led to the emergence of non-proprietary, non-market forms of information production, e.g. Wikipedia, open source software, that harness not only the opinions of the crowd, but also its creative potential. In summary Benkler's argument is concerned not only with the economics of non-proprietary, non-market, information production as an organizational form; but also its implications for law, politics and culture. In contrast to Porat, Benkler places emphasis not only on production activities and processes, but also on the values and consequences of proprietary and non-proprietary production of information, as a primary good in its own right.

Information as an Asset

From an information science perspective, a series of articles by Oppenheim, Stenson and Wilson (2003a, 2003b, 2004) provides the most considered review of the notion of information as an asset. Oppenheim (2003a) identifies two schools of thought relevant to understanding the notion. The first school dates back to the early 1980s where the notion emerges within the context of information resource management (IRM); and the other school is concerned with accounting, and where the notion of information as asset emerges within the context of estimating the value of intangible assets. From an IRM perspective the notion is principally used as a way of identifying and documenting *existing* information assets, with a view to establishing their management and value to the organization. The definition of an information asset drawn up by the Hawley Committee is typical of the IRM approach. For example Horne (1995) makes the following remarks in respect of the emerging importance of information and its governance: "the…common thread is 'information' — its use,

presentation, processing and so on, for the good or otherwise of an organization. The first connection governance is about how an organization uses its assets. If this is put with the second connection — information — the subject being considered is the 'governance of information' or, in other words, the treatment of information as an asset" (Horne, 1995: 6). This IRM has continued via information mapping (Horton, 1988) and by extension inventorying and information flow analysis as part of some approaches to conducting information audits (e.g. Henczel, 2001; Orna, 1999). While it can be argued that the IRM approach is centrally concerned with the role of information resources and their contribution to organizational goals, from an information management perspective the use of IRM has in practice been more concerned with the identification of information and its attributes, e.g. quality, accuracy, timeliness etc. *qua* information assets; rather than establishing their cost, economic value, and contribution to economic productivity and goals. In contrast the accounting school has sought to estimate the value of information as an intangible asset not only in the present but also in the future. From an accounting perspective the notion of information as an asset involves estimates the 'rights or other access to future economic benefits controlled by an entity as a result of past transactions or events" (Oppenheim et al, 2003a, 164). The two schools of thought can be combined into a composite definition, with the authors proposing that "information assets comprise resources that are, or should be, documented and which promise future economic benefit(s)" (Oppenheim et al., 2003: 165). The evidence from Oppenheim's study with both business executives and with information professionals, is that, beyond their identification and mapping, information assets are considered to have value into the future, but that this value tends to be construed in terms of supporting organizational effectiveness, via sense-making and informed decision-making for example, rather than economically. In short both the IRM and intangible asset perspectives draw on the notion of information as an asset. The former is more concerned with determining the cost of an information resource management function, and evaluating this cost against the benefits that that function brings to an organization. The latter is more concerned with the economic value of information assets, now and in the future; and is an approach that has much less currency in information science, and information management.

In summary information work has been approached and understood in a number of ways: as an occupation, as a set of resources activities and processes involved in the production of information or other goods, or an economic asset. Attention is now turned to some developing approaches to data work, the differences between data work and information work, and the implications for data professionals.

Developing Approaches to Data Work

Undoubtedly it is the phenomenon of big data which is the current driver for the emergence of data as an economic asset and societal resource. This phenomenon has led to the development of systematic frameworks for how to create value from data, and how to govern decisions around its capture, quality etc. There is also a recognition that data-related jobs have emerged as an occupational category in their own right. Some of these developments are reviewed here before turning our attention to the challenges that the shift in value from information to data poses for data professionals.

That data work is an emerging occupational category is clearly illustrated by the European Commission's *Digital Agenda for Europe* (https://ec.europa.eu/digital-agenda/en), which incorporates within it a specific focus on the digital economy. Under this broad umbrella a number of surveys have been conducted scoping the extent of the emerging market for data scientists. In the UK for example it has recently been estimated that there has been a "tenfold increase in demand for big data staff in the past five years, with vacancies rising from 1,800 in 2008 to 21,400 in 2013 – an average annual increase of 212 per cent". The biggest demand is for developers (accounting for 41 per cent of advertised vacancies), followed by architects (10 per cent), consultants (10 per cent), analysts (7 per cent), administrators (5 per cent) and data scientists (2 per cent). The demand for these data-oriented jobs is now outstripping the demand for IT and data warehouse or business intelligence staff. The demand for technical skills remains high with applicants likely to need experience in big data (28 per cent), business intelligence (24 per cent), data warehousing (16 per cent), extract transform and load (13 per cent) and analytics (13 per cent). However companies are also looking for business acumen, interpersonal and managerial skills, plus domain knowledge to be able to apply big data insight and transform it into business strategy and action (SAS/Tech partnership, 2015a). A further report points

beyond data specialists, and the need for organizations to build and develop multi-faceted teams with the complementary skills needed to realize the full value of big data: "…it's almost impossible to find one individual with all the technical and soft skills, such as communication and presentation skills, being demanded by business. What's needed in many cases is development of a data science team comprising people with complementary skills" (SAS/Tech partnership, 2015b).

From an organizational perspective two developments around the emergence of data work are noteworthy: the concept of a data value chain; and practices of data and information governance. The concept of a data value chain is already embedded in EU discourse in the form of the development of a European 'data ecosystem': "The current fragmentation of the European data economy and the lack of a thriving European data ecosystem hinder the full exploitation of the enormous economic potential of data to the benefit of European economy and society. A well-functioning data ecosystem is supposed to bring together data owners, data analytics companies, skilled data professionals, cloud service providers, companies from the user industries, venture capitalists, entrepreneurs, research institutes, and universities. In order to support the emergence of a European data ecosystem, a set of regulatory and non-regulatory framework conditions need to be put in place. The issue of fragmentation regarding data and the data value chain with the EU institutions and agencies also needs to be addressed (http://ec.europa.eu/dgs/connect/en/content/data-value-chain-european-strategy). While this policy statement clearly points to the need to develop data value chains at the industry level, the concept is also pertinent at the organizational level. Any organization, whether it be a business, social media company or government organization etc., will want to consider the development of a data value chain which systematically identifies the constituent value-adding processes that turn data in action. Fig 1. Presents one model of this set of linked processes.

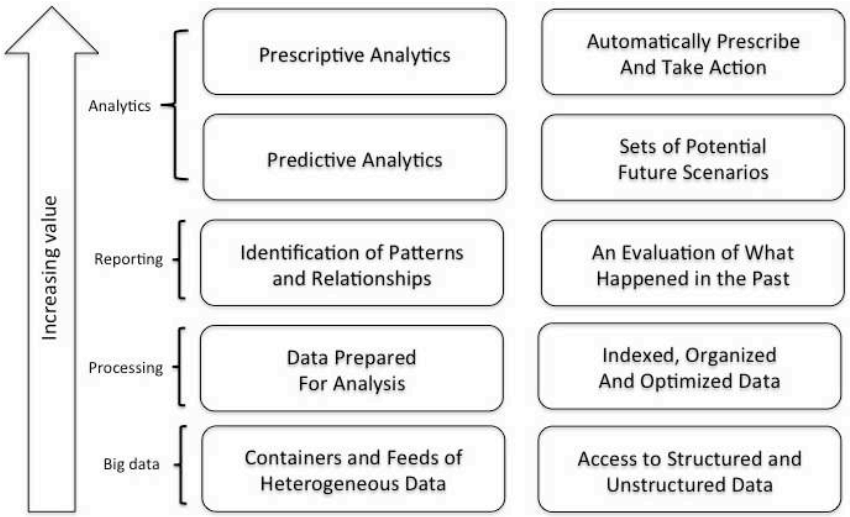

Figure 1. *Analytics Value Chain* (Stein, 2012)

Based on Michael Porter's notion of a value chain (Porter, 1985) and the economic value that cumulatively accrues at each stage of the chain, the model identifies the capture of big data, processing, reporting, and analytics as the key constituent activities. The anchor link is big data, consisting of large varied containers, and sometimes real-time feeds, from heterogeneous sources and types of data, e.g. text, numbers, images, audio, video. Once accessed, the processing link consists of a number of actions that can be applied to the data, e.g. assigning metadata, face-recognition, which are preparatory to subsequent analysis. The aim of the reporting link is to identify patterns and relationships in the prepared data, and to present and visualize these for example via dashboards. It is then the goal of the analytics phase to interpret, make sense of, or otherwise take decisions on the basis of the data presented. This can involve decisions and actions geared to the enhancement of existing products or services, or the development of new products and services.

Each of these links in the chain has been and will continue to be the focus of specific attention (Cukier and Mayer-Schonberger, 2013; Few, 2006; Davenport and Harris, 2010). From the perspective of this article, this attention includes an emerging literature on the governance practices related to the processing of data as an economic asset (Davenport, 2014; Khatri and Brown, 2010). Khatri and Brown (2010) for example define data governance in reference to "who holds the decision rights and is held accountable for an organization's decision-making about data assets", and

propose five decision domains which are required to maximize the value to be derived from data assets: data principles, data quality, metadata, data access and data life-cycle. Data principles "set the boundary requirements for the intended uses of data" and are needed to clarify 'the role of data as an asset'. Decisions around data quality are needed to establish the requirements surrounding the 'intended use of the data'. Decisions around metadata are needed to establish the "semantics or "content" of data so that it is interpretable by the users"; decisions around data access are needed for "specifying access requirements of data", while decisions around the data lifecycle are needed for "determining the definition, production, retention, and retirement of data" (Khatri and Brown, 2010: 149). With all these decisions there will be a tension between the extent to which the locus of accountability for the decisions is centralized or decentralized. Of the many merits of Khatri and Brown's approach one is that it builds on an established framework already used for for IT governance (Weill and Ross, 2004); while at the same time introducing an element of accountability and stewardship in relation to data. However it should be added that the kind of accountability that the authors refer to is one motivated by the business and organizational value of the data, rather than accountability to an external stakeholder, e.g. a consumer, client, citizen, or legal and regulatory frameworks.

It is clear that the implementation of data value chains, and within this practices of data governance, point to the increasing value that organizations are placing on data as an economic asset. At the same time the emergence of 'information governance' has sought to develop a set of practices that seek not only to exploit the value of data, but also to mitigate the risks that an increased emphasis on the development of data-intensive products and services places on organizations. This is the case principally because much of the data work involves the processing, and analysis of personal identifying information (PII).

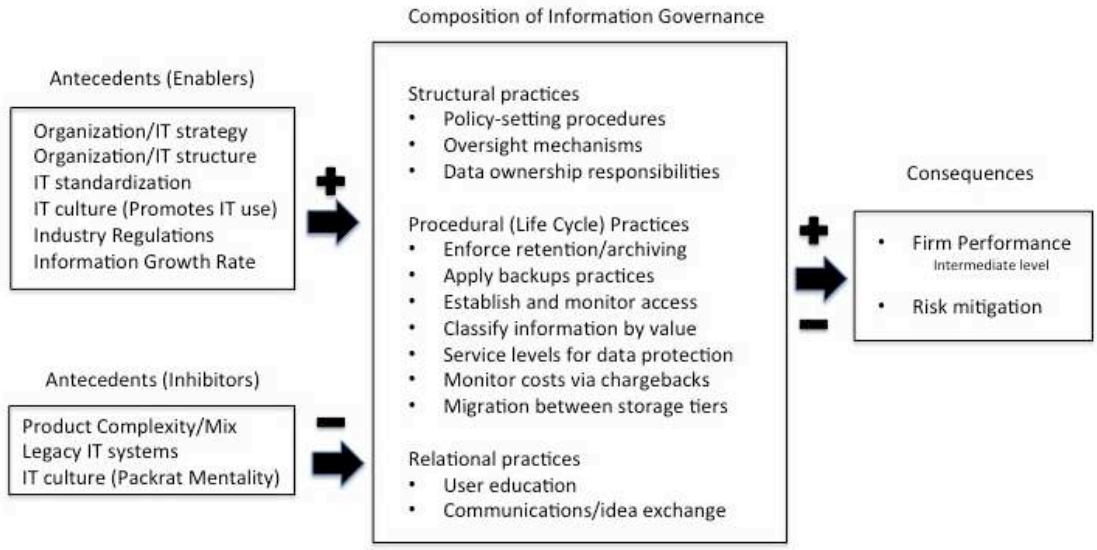

Figure 2. *Information Governance Research Model*
(Tallon, Ramirez and Short, 2013)

Motivated by firm performance and risk mitigation, information governance can be defined as "a collection of capabilities or practices for the creation, capture, valuation, storage, usage, control, access, archival, and deletion of information over its life-cycle" (see Fig. 2) (Tallon, Ramirez and Short, 2013: 142). These practices comprise of procedural practices, structural practices, and relational practices. Procedural practices are the "mechanical arm of information governance [and] are both technical and managerial. Technical practices describe how systems have automated migration of data between tiers or how additional storage resources are provisioned, or how systems are used to govern access and backups. Managerial practices describe how data is classified (data classification) so that storage decisions can be made based on differential value characteristics" (Tallon, Ramirez and Short, 2013: 164-165); structural practices are associated with "setting the locus of IT decision making or data stewardship; assignments of roles to key decision makers; practices associated with IT reporting structure; use of oversight committees or other high-level policy setting/monitoring groups" (Tallon, Ramirez and Short, 2013: 156); and relational practices "show how organizations build knowledge among users around the need for information governance and how they work with conflicting policies" (Tallon, Ramirez and Short, 2013: 165) and will include a focus on education, knowledge sharing, and conflict resolution. Tallon, Ramirez and Short (2013) also point to factors in the organizational environment that can influence the composition and

implementation of information governance. Depending on their current state, an organization's IT culture or IT infrastructure can either enable or inhibit the implementation of information governance.

In summary, in response to the emergence of data as an economic and societal resource, current approaches to data work are revolving around the development of data value chains, their constituent processes, and information governance and decision-making practices. All of these approaches focus primarily on the organizational context ,and do not effectively address how data professionals should be aware of the broader economic and societal conditions impinging on the emergence of data work, plus the consequences of organized data work for organizations and users. The final section addresses some of these emerging challenges.

Understanding Data Work in Context: From Data Scientists to Data Professionals

Fig. 3 sets out how data work can be more effectively approached by placing it in the context of the broader informational, organizational, sectoral, national, and international conditions and consequences related to data work. It is important to note that the 'matrix' illustrates yet to be discovered connections between the different levels or contexts, which are pertinent to understanding a phenomenon. Building on Tallon, Ramirez and Short (2013) and grounded theory (Corbin and Strauss, 2008), the matrix can be best understood as a series of interconnected circles with arrows pointing towards and away from the process of interest. In this instance the key process is that of data work, with the arrows pointing to that process acting as the conditions influencing the conduct of data work and how it is done, with the arrows pointing away from the centre acting as the consequences of doing data work. It is important to recognise that the diagram is schematic only and is intended to act as a tool for designing the research training programme, its individual projects, and complementary skills. It not intended to rigidly point to existing connections between the different levels since an understanding will only emerge from the research undertaken. Beginning at the outer edge of the circles, the broadest macro area is that of the *international* context of a global economy driven by technology adoption, but also influenced by international relations and regulations. Both of these aspects will

involve politics between and within states not the least of which will concern the mobility, training, and conditions of an international workforce. It is this international context that we can place the EU digital economy and society, the competitiveness of its digital economy in relation to other nations and the development of its own. The next circle in identifies the *national* context, within which can also be placed values that have broader scope and applicability beyond distinct sectors of the economy and of society.

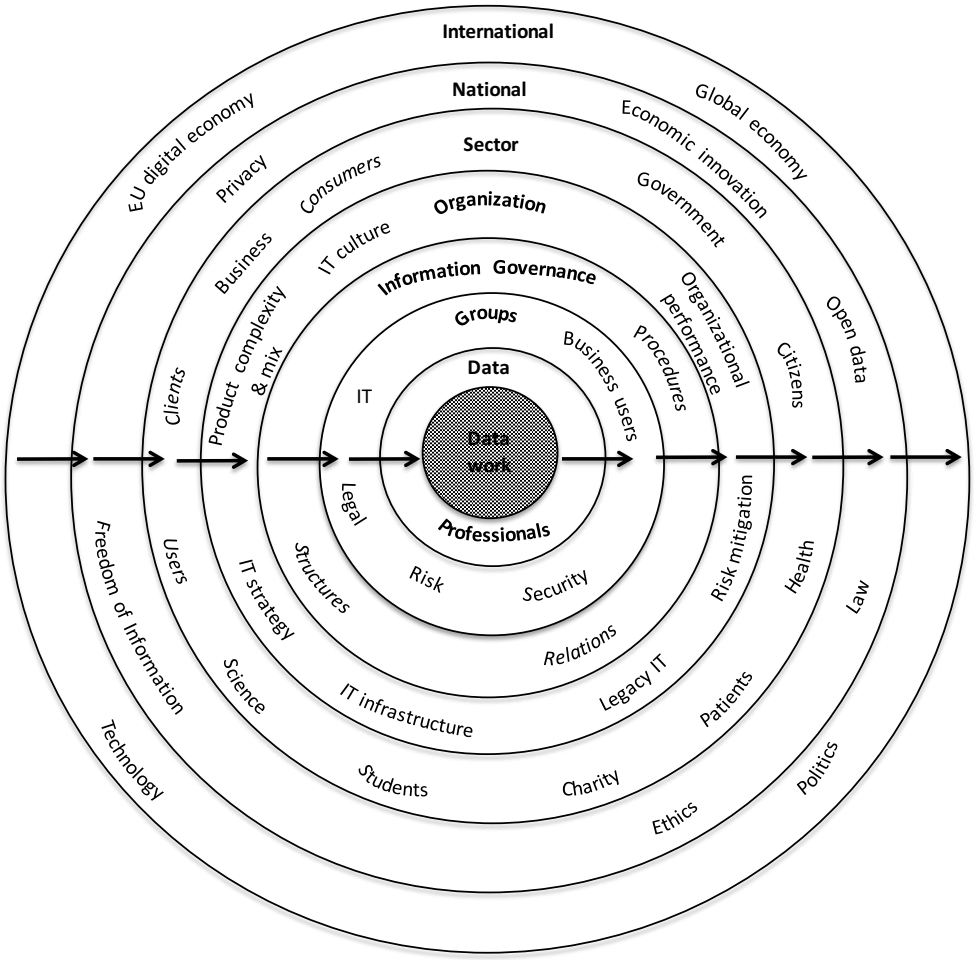

Figure 3. *Data Work in Context*

Therefore we include in this circle on the one hand national economies and innovation and on the other law, legal regulations; but also ethics, freedom of information and privacy as values upheld and practiced in varying degrees by individual organizations and individual members within a country. Exploration of both these international and national contexts goes some way to partially answering the question as to why data

work is been carried out i.e. for economic gain but in a context of existing regulations, and values. The next two circles in identify where data work takes place i.e in specific *sectors* e.g. business, health, charities, and science, and the *organizations* located within each of these sectors. More specifically the *organization* circle identifies the context and conditions at the organizational level influencing why, what and how data work is carried out. These include on the one hand improving organizational performance, but also mitigating the risks, e.g. economic, legal, reputational of doing data work; and culture. At the same time there will be a number of organizational factors, that can either enable or constrain data work processes. These principally include different aspects of an organization's IT capability, including IT infrastructure, legacy systems, and culture, along with the complexities of what the organization is attempting to produce or deliver. Within organizations there exists an information environment, the main aspect of which concerns the presence or absence of a number of information governance and other work practices that influence the data work environment e.g. data governance procedures for assuring the relevance and quality of any data captured, processed, and used; roles and responsibilities of data stewards and others who structure how the data work environment is organized; along with training, collaboration and the establishment and maintenance of working relations between data professionals and other specialists. The next circle identifies who these specialists are including IT professionals, legal specialists, risk and security professionals, and business users involved in taking product, service and other decisions on the basis of the data processed. All of these circles and contexts identify conditions that influence the data professionals and data work that sits at the core of the diagram. Organized data work can be conceived of as a value-adding process beginning with data, and incorporating distinct phases including processing, reporting and analytics phases. Viewed in this way, it is also reminiscent of Taylor's (1986) value-added spectrum of data, information, knowledge and action (see Fig.4). Drawing on this spectrum, data work can be conceived of as a set of activities that begins with data, and the organization of data into information, continues with the analysis of information, and its turning via interpretation and other processes into knowledge. Judgment is then applied to the knowledge, which is then made productive via practical decision-making, with the spectrum completed with action and use. In this sense, data work can be conceived of as incorporating data,

information, knowledge, and action as part of the same value-added spectrum. Therefore, although incorporating information, it is nevertheless convenient to call it data work, since the purpose of the process is to make data actionable. This is different from information work, which is centrally concerned with the access, processing, and use of already encoded data.

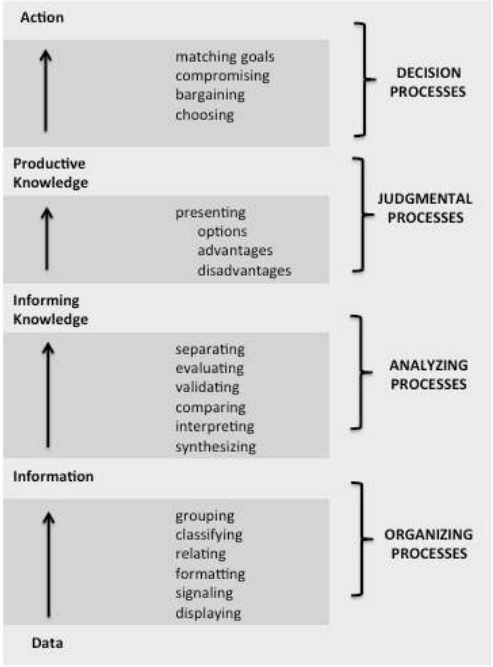

Figure 4. *Value-Added Spectrum* (Taylor, 1986)

The arrows leading away from data work and data professionals point to the consequences of doing so. In other words, the increasing interest in the value of data as both an economic asset and a societal good, and its conditions, will have consequences at other levels of the diagram. For example, an increase in data work will have an influence on recruitment of data professionals, or the emergence of data-intensive organizations will have consequences for citizens, consumers, patients etc.; in turn there will be an impact on legal regulations and ethical considerations. Beyond this the contribution of data-driven products and services will have an impact on the size and nature of digital economy. This brings us to a final important principle of the diagram; that data work and its processes are not only conditioned by and have consequences for individuals, organizations, and economies - these consequences influence the initial conditions. Therefore, we can speak of a matrix of interrelated conditions and consequences at macro, meso and micro levels.

Placing data work in such a context also serves to better illuminate the governance and education challenges relevant to organizational data work. And it is this context and the issues that arise which is atypical of the current education of the data scientist and yet their consideration and prevention is critical to more human-centred data work. Take this definition of the data scientist developed by the NIST Big Data Working Group. A data scientist is "someone who has sufficient knowledge in the overlapping regimes of expertise in business needs, domain knowledge, analytical skills, and programming and systems engineering expertise to manage the end-to-end scientific method process" (Demchenko 2015). At no point in this definition do we have an indication of the broader issues of accountability, legal and regulatory frameworks and ethics.

While the use of data for economic purposes has always existed in tension with the privacy of the users of organizations' products and services, the emergence of a digital economy has contributed to intensifying the problem. Since, in a digital economy, organizations not only passively capture and process data about users, but users also actively add value to the platforms and services that they use. Users do this by contributing opinions and other information to systems e.g. Amazon's Customer Review system, Eopinions, IMDb, TripAdvisor; by valorising the production and distribution of free versions of digital content, via initial contact, attention and its subsequent monetization e.g. Google, Facebook; and by engaging in the informational co-production of goods via commons-based peer production e.g. archives, Wikipedia, open source (Foster, 2013; Foster, Benford and Price, 2013). While the subsequent human and automatic processing of these valued contributions raises considerations of privacy, consent, and the security of personal data in a digital economy, Dormehl (2015) also points to other consequences – both intended and unintended – of pervasive data capture and the algorithmic processes at work in our society, including mis-categorization, and inaccurate profiling. In summary data work is much more tied to context than either information work, or more technicist approaches to data work grant. This is by virtue both of the personal identifying information that is being processed e.g. location-based data and the uses to which data-driven products and services are put e.g. personalization. The matrix can aid not only in developing a more human-oriented and sustainable approach to data work, one that places data work in

the context of its conditions and consequences; but it can also serve as a map on which to locate a number of different starting-points for practical research.

References

Bell, D. (1973). *The Coming of Post-Industrial Society*: *A Venture in Social Forecasting*. New York: Basic Books.

Benkler, Y. (2006). *The Wealth of Networks*: *How Social Production Transforms Markets and Freedom.* New Haven: Yale University Press.

Choo, C-W. (2001). *Information Management for the Intelligent Organization*: *The Art of Scanning the Environment*, Third edition, Medford (NJ): Information Today.

Corbin, J. and Strauss, A. (2008). *Basics of Qualitative Research*: *Techniques for Developing Grounded Theory*, Third edition, London: Sage.

Cukier, K. and Mayer-Schonberger, V. (2013). *Big Data: A Revolution That Will Transform How we Live, Work and Think.* London: John Murray.

Davenport, T.H. (2014). *Bigdata@work: dispelling the myths, uncovering the opportunities*. Boston (Ma.): Harvard Business Review Press.

Davenport, T.H. (2010). *Analytics at Work: Smarter Decisions, Better Results*. Boston (Ma.): Harvard Business School Publishing.

Demchenko, Y. (2015). EDISON: Coordination and cooperation to establish new profession of data scientist for European research and industry. Available at: *https://rd-alliance.org/.../OpenScholarshipEdison_YuriDemchenko.pdf*

Dormehl, L. (2015). *The Formula*: *How Algorithms Solve all our Problems ... and Create More.* London: WH Allen.

Few, S. (2006). *Information Dashboard Design*: *The Effective Visual Communication of Data*. Sebastopol (Ca.): O'Reilly.

Foster, J. (2013). "Valorising the cultural content of the commodity: On immaterial labour and new forms of informational work". In: Lin, A., Foster, J., and Scifleet, P. *Consumer Information Systems and Relationship Management*: *Design, Implementation and Use*, 189-201. Hershey, PA: IGI Global.

Foster, J., Benford, S. and Price, D. (2013). "Digital archiving as information production: using experts and learners in the design of subject access", *Journal of Documentation*, **69** (6), 773-785.

Hardt, M. and Negri, A. (2000). *Empire*. London: Harvard University Press.

Hawley Committee. (1995). *Information as an Asset The Board Agenda: A Consultative Report*. Hawley Committee/IMPACT programme.

Henczel, S. (2001). *The Information Audit: A Practical Guide*. Munich: K.G. Saur.

Horne, N.W. (1995). "Information as an Asset The Board Agenda", *Computer Audit Update*, September, pp. 5-11.

Horton, F.W. (1988). "Mapping Corporate Information Resources", *International Journal of Information Management*, **8**, 249-254.

Khatri, V. and Brown, C.V. (2010). "Designing data governance", *Communications of the ACM*, 53 (1), 148-152.

Laney, D. (2001). "3D data management: Controlling data volume, velocity, and variety". Available at: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Oppenheim, C., Stenson, J. and Wilson, R.M.S. (2003a). "Studies on information as an asset I: definitions", *Journal of Information Science*, **29** (3), 159-166.

Oppenheim, C., Stenson, J. and Wilson, R.M.S. (2003b). "Studies on information as an asset II: repertory grid", *Journal of Information Science*, **29** (5), 419-432.

Oppenheim, C., Stenson, J. and Wilson, R.M.S. (2004). "Studies on information as an asset III: views of information professionals". *Journal of Information Science* **30** (2), 181-190.

Orna, E. (1999). *Practical Information Policies*. Aldershot: Gower.

Porat, M.U. (1977). *The Information Economy*: *Definition and Measurement*. Washington (DC): Department of Commerce/Office of Telecommunications.

Porter, M.E. (1985). *Competitive Advantage: Creating and Sustaining Superior Performance.* London: Free Press.

Reich, R.B. (1991). *The Work of Nations*: *Preparing Ourselves for 21st Century Capitalism*. Alfred Knopf: New York.

SAS/Tech Partnership (2015a). *Big Data Analytics: Assessment of Demand for Labour and Skills 2013-2020*. London: SAS/Tech Partnership.

SAS/Tech Partnership (2015b). *What Makes a Great Data Scientist*: *Survey Summary Report.* London: SAS/Tech Partnership. Available at: http://www.sas.com/en_gb/offers/14q4/data-scientist-report.html

Smallwood, R.F. (2014). *Information Governance*: *Concepts, Strategies, and Best Practices*. Hoboken (HJ): Wiley.

Stein, A. (2012). "Big data and analytics, the analytics value chain – part 3". Available at: http://steinvox.com/blog/big-data-and-analytics-the-analytics-value-chain/

Tallon, P., Ramirez, R.V. and Short, J.E. (2013). "The information artifact in IT governance: Toward a theory of information governance", *Journal of Management Information Systems*, 30 (3), 145-181.

Taylor, R.S. (1986). *Value-Added Processes in Information Systems*. Norwood (NJ): Ablex Publishing Corporation.

Webster, F. (2006). "What is an information society?" In: *Theories of the Information Society*, pp. 8-31, Third edition. London: Routledge.

Weill, P. and Ross, J.W. (2004). *IT Governance: How Top Performers Manage IT Decision Rights for Superior Performance*. Boston (Ma.): Harvard Business School Publishing.

Wilson, R.M.S. and Stenson, J.A. (2008). "Valuation of information assets on the balance sheet: the recognition and approaches to the valuation of intangible assets", *Business Information Review*, **25**, 183-189.