BIG DATA & COMPLEX KNOWLEDGE

OBSERVATIONS AND RECOMMENDATIONS FOR RESEARCH FROM THE KNOWLEDGE COMPLEXITY PROJECT

AUTHORS:

Jennifer Edmond, Nicola Horsley, Elisabeth Huber, Rihards Kalnins, Joerg Lehman, Georgina Nugent-Folan, Mike Priddy, Thomas Stodulka

The Knowledge Complexity (KPLEX) Project
was funded by the European Commission Grant
Agreement number 732340, 2017-2018.

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

TABLE OF CONTENTS

I. INTRODUCTION: BIG DATA AND THE KNOWLEDGE COMPLEXITY PROJECT 5

II. WHO IS THIS DOCUMENT FOR? 6

III. INTEGRATED FINDINGS AND AREAS FOR FUTURE WORK 8

A. Big data is ill-suited to representing complexity: the urge toward easy interrogability can often result in obscurity and user disempowerment. 8

B. Big data compromises rich information 9

C. Standards are both useful and harmful. 9

D. The appearance of openness can be misleading. 10

E. Research based on big data is overly opportunistic. 11

F. How we talk about big data matters. 12

G. Big data research should be supported by a greater diversity in approaches. 13

H. Even big data research is about narrative, which has implications for how we should observe its objectivity or truth value. 14

I. The dark side of context: dark linking and de-anonymisation. 15

J. Organisational and professional practices. 15

K. Big data research and social confidence. 16

IV. RECOMMENDATIONS 17

V. IMPACT OF THE KPLEX PROJECT 21

I. INTRODUCTION: BIG DATA AND THE KNOWLEDGE COMPLEXITY PROJECT

The Knowledge Complexity (or KPLEX) project was created with a two-fold purpose: first, to expose potential areas of bias in big data research, and second, to do so using methods and challenges coming from a research community that has been relatively resistant to big data, namely the arts and humanities. The project's founding supposition was that there are practical and cultural reasons why humanities research resists datafication, a process generally understood as the substitution of original state research objects and processes for digital, quantified or otherwise more structured streams of information. The project's further assumption was that these very reasons for resistance could be instructive for the critical observation of big data research and innovation as a whole. To understand clearly the features of humanistic and cultural data, approaches, methodologies, institutions and challenges is to see the fault lines where datafication and algorithmic parsing may fail to deliver on what they promise, or may hide the very insight they propose to expose. As such, the aim of the KPLEX project has been, from the outset, to pinpoint areas where different research communities' understanding of what the creation of knowledge is and should be diverge, and, from this unique perspective, propose where further work can and should be done.

The KPLEX project team was recruited in such a way as to be create an experiment and a case study in interdisciplinary, applied research with a foundation in the humanities. Each of the four partner research groups was drawn from a very different research community, with different fundamental expectations of and from the knowledge creation process. The team of four partners included research groups in both digital humanities and anthropology, a research data archive and an SME specialising in language technologies. This diversity was a strength of the project, but also a constant reminder of how challenging such cooperative work, across disciplines and sectors, can be.

Although the KPLEX project had only a short duration (15 months), its results point toward a number of central issues and possible development avenues for a future of big data research that is socially aware and informed, but which also harnesses opportunities to explore new pathways to technical innovations. The challenges for the future of this research and for its exploitation will be to overcome the social and cultural barriers between the languages and practices not only of research communities, but also of the ICT industry and policy sectors. The KPLEX results point toward clear potential value in these areas, for the uptake of the results, their application to meet societal challenges, and for improving public knowledge and action. Such reuse, however, may take significant investment and time, so as to establish common vocabulary and overturn long-standing biases and power dynamics, as will be described below. The potential benefits, however, could be great, in terms of technical, social and cultural innovation.

II. WHO IS THIS DOCUMENT FOR?

Given the broad aims and objectives of the project as defined in above, the results and the example of the KPLEX project are of use to a wide variety of potential audiences.

For **researchers**, the example of KPLEX has documented how high the impact can be within a broadly interdisciplinary project, looking at technology by drawing upon the perspectives of literary analysis, anthropology, library science and others. The techniques by which we have both differentiated and aligned our standpoints stands as a case study in the integration of approaches and data across a number of potential fault lines. Research aiming to build upon KPLEX's results will also be smoothly facilitated by the project's open sharing of its research data.

For **policymakers**, KPLEX has achieved its primary aim of creating an empirical basis that exposes sources of bias in big data research. Its results show, in an integrated and holistic way, what issues might form a focus for future work, and what fissures might be approached, via regulatory, policy or practice interventions, to improve upon the current situation. Each of our thematic cases was defined to address a specific policy requirement currently visible at European level: the need for more responsible approaches to funding the development of big data; the need to increase the possibility that cultural data can be shared and reused effectively; the need to broaden the possible pool of knowledge available to research and industry through the fostering of open science; and the need to contribute new insight to culturally sensitive communication tasks, as in multilingual environments. Work in each of these areas will be able to draw from the KPLEX results.

For the **ICT industry and research**, KPLEX's results may at times seem challenging, but this alternative perspective should become a source of inspiration, rather than frustration. Software development exists in many ways as a distinct culture, with its own language, norms, values and hierarchies. As with any culture, these norms and values can provide a strong platform for creativity and development, but can also prove a hindrance in situations that require translation and negotiation with another such 'culture'. At a time when the need for privacy preservation and a stronger ethical focus are becoming ever more widely recognised in ICT development and regulation, KPLEX's results should be a welcome source of fresh thinking. They can encourage deeper probing into fundamental areas of research, such as managing uncertainty, supporting identity development, or exploring the unexpected impacts of digital interventions in society, all of which may be taken for granted in an innovation monoculture.

KPLEX has developed a strong resonance with **citizens**, having attracted a number of high-profile national broadcasters to feature the project. This is a reflection not only of the quality and accessibility of the project, but also of the Zeitgeist in which it has been developed. People know enough about big data research to be concerned by it, but the inter-disciplinarity of KPLEX has made it a very fertile ground for public outreach, given that as multidisciplinary researchers in the KPLEX team, we cannot ourselves retreat to disciplinary networks and jargon to communicate our results.

Finally, KPLEX has uncovered significant patterns in the **organisational and institutional** responses to the rapid changes being brought about by ICT, the gaps that are being left and the opportunities that are being found. Amongst researchers and practitioners alike, KPLEX has discovered that the potential good of big data research is shadowed by real and justifiable feelings of knowledge and perspectives being left behind, of being overwhelmed, of a loss of control and of diminishing authority for long-established practices and their underappreciated functions. As such, the project findings will be of use and interest to organisations struggling with technology adoption in the face of rapid change on the one side and no decrease in the importance (and resource intensity) of their pre-digital missions on the other.

III. INTEGRATED FINDINGS AND AREAS FOR FUTURE WORK

The KPLEX project was conceived of and organised according to a set of four themes: discourses of data, hidden data, human bias in data and the loss of cultural information in data. In the course of researching these themes, KPLEX mined the attitudes and opinions of many different researchers and professionals, through literature reviews, interviews, surveys and exploration of other material, such as scientific articles. Each thematic strand produced insightful and significant results, but the most compelling outcomes of the project stand at the intersection these themes and cohorts. The resonances between and across the perspectives mined by the project illuminate areas where we can evidence fundamental challenges to big data research, or opportunities for innovative future activities. These topics will not be simple to pursue, since some of them (as the discussion below will explain) are viewed by key contributors as unnecessary barriers to technical progress. It is clear, however, that such inconvenient truths of big data research are beginning to have an undesirable societal impact, and the KPLEX conclusions, while requiring courage to implement, can provide a solid foundation point for addressing many of them.

A. BIG DATA IS ILL-SUITED TO REPRESENTING COMPLEXITY: THE URGE TOWARD EASY INTERROGABILITY CAN OFTEN RESULT IN OBSCURITY AND USER DISEMPOWERMENT

The fulfilment of the technical need to render complex phenomena in a binary system of 1s and 0s feeds into a very human attraction to answers that are simple, straightforward, confident, and possibly even false, or at least misleading. Big data researchers tend to portray complexity as a negative, rather than a positive, which commits the research area to the marginalisation, removal, structuring or 'cleaning' of complexity out of data. The KPLEX project was able to observe the resistance among many information experts and researchers to such simplifications. In particular emotion researchers, who look at very complex, and often contradictory, human phenomena, were able to express elegantly those aspects of their research they would not be able to capture and quantify as data, such as identity, culture and individual emotions. The fact that such signals 'operate below conscious awareness in their actual practice' and that 'people can't always access and articulate their emotions' is therefore a great example of the kind of challenges inherent when we try to represent human activity in the form of data.

B. BIG DATA COMPROMISES RICH INFORMATION

One of the most common recurrent themes across the KPLEX project interviews was that of how big data approaches to knowledge creation both lose and create context. Context can encompass a huge range of indicators of how data can and should be reused, such as its provenance, how it came to be created, and the humans and biases that may lurk behind its collection or creation.

Cultural heritage professionals and researchers alike recognised the potential implications of stripping away too much in the datafication process. Catalogue records in libraries and archives were viewed with some suspicion, for example, in recognition of the fact that they were not meant to be used in isolation from the tacit knowledge of the professionals who create and preserve such records. This may be the reason that researchers studying emotion by and large eschew the use of the existing standardised description languages in their descriptions and analyses.

Similarly, keyword searches, such as are widely facilitated by popular search engines, also represent a form of impoverishment, a single strong channel for knowledge discovery that eclipses a large number of other powerful but more subtle ones. A feeling of getting to know material, of a discovery process approaching intimacy, is bypassed by this approach, specifically because of the layers of context it strips away. As one interviewee stated it, ‘when you go with the direct way, in the current state of the search engines, you miss the information.’ The problem, of course, is that the potential for context has no boundaries, and no description can ever be said to be fully complete. Professional archivists therefore take the need to meet the optimal compromise in capturing context in order to support the appropriate use of their holdings as one of their most important duties and greatest challenges. This is what they are trained to do, but it is both an art and a craft, and one that is not always valued in a system where the finding aid is perhaps only ever ‘seen’ by an algorithm.

C. STANDARDS ARE BOTH USEFUL AND HARMFUL

Many approaches to data management that are considered as ‘standards’ are looked upon as suspicious or indeed destructive to knowledge creation by researchers, and indeed by knowledge management professionals (such as librarians and archivists). Such commonly accepted big data research processes as data cleaning or scrubbing were often characterised as manipulations that have no place in a responsibly delivered research process or project. Researchers and professionals who work with human subjects and cultural data express a strong warning that we should not forget that there is no such thing as 'raw' data: the production of data is always the product of someone's methodology and an epistemology, and bears the marks of their perspective, in particular where the phenomena described in the data are complex and/or derived from individual experience. If KPLEX has proven anything, it is that knowledge creation professionals in areas that draw upon the messy data produced by human subjects are suspicious of big data for the manner in which it discards complexity and context for the sake of technical processability. This transformation process, also known as 'datafication,' from the lived to the digital, and from the complex to the computable, is understood as necessarily and implicitly a loss of information, be that sensory, tacit, unrecognised, temporally determined or otherwise susceptible to misrepresentation or non-representation by digital surrogates. It is useful to note that the 'noise' removed in the pursuit of the 'signals' in this process is often not documented or preserved. To go even further, the creation of data sources, such as archival descriptions or interview transcripts, is clearly perceived as interpreted the expression of a power dynamic.

Information loss may occur at any stage of a datafication process, but undergoing classification probably has the most lasting effects. The a priori relegation of a phenomenon into distinct categories, like for example the reduction of a person's wide array of affective experiences and feelings into a small number of basic emotions (like happiness or anger) clearly restricts knowledge, and can potentially mislead. Rigid classification schemes not only have consequences on scientific research, but also shape public discourse: when they are too rigid or too reductionist, they can have social consequences, and are hence political.

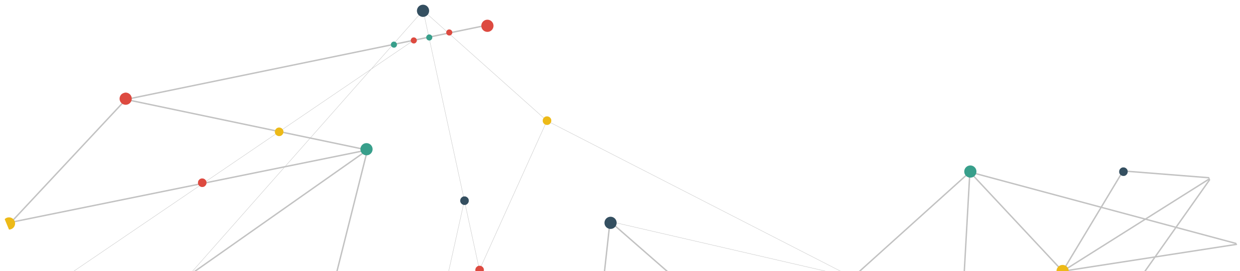
D. THE APPEARANCE OF OPENNESS CAN BE MISLEADING

The fact that some of the best known, consumer-facing big data industry leaders, such as Google, Facebook or Twitter, operate under a business model that provides services to the user for free (though this of course can be debated) leads to the perception that such platforms are open, democratic, and unbiased. Against this simple perception, however, such platforms were consistently referred to in the KPLEX interviews as representing a threat to access and to the development of unbiased knowledge. On the one hand, this perception is based upon the recognition that the data such platforms captures and holds is a corporate asset and basis for corporate profit, albeit one based on the contributions of many private individuals. On the other hand, the network effect of such all-encompassing platforms creates dominant forms of information retrieval and knowledge production that, in spite of their inherent biases and limitations, may be gradually eclipsing other, potentially complementary, potentially more powerful, equivalents. A Google search may indeed be faster than a consultation with an archivist, but it only draws on one form of record, explicit and electronic, it is potentially without verification, and may even be intended to mislead.

The digital record can suffer from impoverishment due to what can be captured explicitly and effectively. As one researcher described it, ‘all this documentation stuff functions as a kind of exogram or external memory storage … the sensual qualities of field notes, photographs or objects from the field have the capacity to trigger implicit memories or the hidden, embodied knowledge.’ Big data systems cannot reflect a tacit dimension, or a negotiated refinement between perspectives: ‘all we access is the expression.’ And not all expressions are created equal. If we are concerned about the development of pan-European identities, for example, and of the strength of cultural ties able to create resilient societies, then we should be very concerned about how the digital record, for all of its global reach and coverage, represents cultures and languages unequally. As one interviewee stated, you have to ‘know what you can’t find.’ If the system appears open, but is in fact closed, your sense of your own blind spots will be dulled, and the spectre of openness will work as a diversion away from both the complex material a system excludes as well as from any awareness of the hiddenness behind the mirage of openness.

E. RESEARCH BASED ON BIG DATA CAN BE OVERLY OPPORTUNISTIC

Interviewees heavily critiqued research founded upon big data for its lack of an ‘underlying theory.’ Rightly or wrongly, they largely viewed big data research as driven by opportunities (that is by the availability of data) rather than by research questions in the conventional sense. According to this conception, data are inseparably linked to the knowledge creation process. More data do not necessarily lead to more insight, nor are big data devoid of limitations, especially with respect to questions of representativeness or bias. It is the algorithms used for the analysis of big data which introduce statistical biases, for example, or which reflect and amplify underlying biases present in the data. The risks inherent in these reversals of the traditional research process include the narrowing of research toward problems and questions easily represented in existing data, or the misapprehension of a well-represented field as one worth investigation.



F. HOW WE TALK ABOUT BIG DATA MATTERS

From the earliest points in human history, we have recognised that words have power. This is still true, and the language used to describe and inscribe big data research is telling. This phenomenon begins, but does not necessarily end, with the term 'data' itself. Among computer science researchers working with big data, including those interviewed by the KPLEX project, this word can refer to both input and output; it can be both raw and highly manipulated. It comes from predictable sources (like sensors) and highly unpredictable ones (like people). Most importantly, it is both yours and mine. The sheer scale and variance of the inconsistencies in definitions appearing the in KPLEX corpus and the variability of what data can be, how it can be spoken of, and what can or cannot be done with it, was striking. The pervasiveness of this super-term is hard to fathom: to give one illustrative example from the project results, in one single computer science research paper, the word data was used more than to 500 times over the course of about 20 pages. This is clearly at the far end of a continuum of use and abuse of the term in question, but the KPLEX researchers observed concerning trends across the discussions of data, including a lack of discrimination between processes or newly captured data, and references to data having such innate properties as being 'real.'

Interestingly, this narrowing of discursive focus in computer science meets explicit resistance in other disciplines. The reluctance among humanities researchers to use the term 'data,' often seen as a sign of their commitment to traditional modes of knowledge creation, goes hand in hand with the reluctance to see research objects as all of one type. Within this cohort, a much richer equivalent vocabulary exists, including 'primary sources,' 'secondary sources,' 'theoretical material,' 'methodological descriptions,' etc. From this perspective, it seems more progressive than regressive that humanists often could not see the data layer in their work, replying instead that they had 'no data to share' or that data was 'not my kind of work.'

Such variations in application of a single word can act as a barrier to reuse of results, to interdisciplinary cooperation, to academic transparency, and to the management of potential social risk. The impact of discourse was interestingly polarising among the KPLEX interviewees, however. Among computer science researchers, such a discussion was perceived as a distraction, as 'anthropomorphised,' impractical, or overly theoretical, philosophical. The telling, but honest, statement of one interviewee about this issue was that 'the computer scientist doesn't care! They just need to have an agreed term.' But the impatience of the computer scientist to move toward a solution is met with a potential ignorance of their methods and discourse on the part of the potential users and subjects of their work: both archivists and researchers reported versions of this conflict, and specifically of using similar words to mean different things, or taking a very long time to find the words in their respective professional vocabularies that meant the same thing. Language is not only about communication however, it is about power, and while we can assume that the language around big data research is not intended to obfuscate or test the authority of the non-ICT proficient to question methods or outcomes, the result may be the same.

G. BIG DATA RESEARCH SHOULD BE SUPPORTED BY A GREATER DIVERSITY IN APPROACHES

Big data research should be a means, but not an end. While computer or data scientists may be able to extract a certain kind of knowledge from large data sets, by their very nature the original sources contain more complexity than those results necessarily represent. Decision-making in big data research should not be driven by perceived technical imperatives to meet an algorithmic challenge or commercial imperatives to serve a market niche, but must also contain a natural braking function to ensure that the technical and the commercial don't outstrip the human and the social. We know that biased data manipulated by biased teams leads to biased software, and we know that abuses of big data 'black boxes' exist: what we do not know is what the opposite of the current imbalance might look like, where truly integrative understanding drives an approach to technical progress.

KPLEX has proven, through its methods and its results, that such mixed teams can generate powerful and actionable insight, but that the success factors for such work have much to do with evening out the engrained power dynamics and facilitating fundamental shared understanding and values, such as an early negotiation of key terminology. Too many interdisciplinary projects proceed, perhaps through their entire life cycles, without ever developing the shared languages required to enable partners to collaborate from a position of parity, not as masters of each others' disciplines and approaches, but as eager observers and students able to understand the first principles and ask the right questions, with confidence and humility, at the borders of their expertise.

Aside from the commonly discussed benefits of interdisciplinary research, such as fostering innovation by convening a mix of approaches and expertise, or checking biases through diversity, further potential strengths can be observed in the KPLEX results. For example, consistency of definitions was more notable among researchers with the same disciplinary training (such as computational linguists), and among researchers who had been working together on the same project or team. Many researchers with experience of interdisciplinary work expressed concern, however, regarding how ‘the other side’ interpreted and worked with ‘their’ data. While some embraced their role as mediators between disciplines, others spoke disparagingly about the respective abilities of engineers or humanities researchers to fully comprehend what they were working with. Growing a culture of greater cooperation between diverse experts will not be simple or straightforward, but the value will be great.

H. EVEN BIG DATA RESEARCH IS ABOUT NARRATIVE, WHICH HAS IMPLICATIONS FOR HOW WE SHOULD THINK ABOUT ITS OBJECTIVITY OR TRUTH VALUE

Big data was highly critiqued for its tendency to remove context, for the manner in which complexity may need to be stripped away to support computability. Context is also about narrative, however. Human beings think in terms of stories, of connections and of relationships between events and information far more than in isolated, unconnected units of information

In the end, even the outputs (e.g. research papers, but also software) of computer science researchers are narrative, not data. To the extent that the word can be said to mean any one thing, data generally seems to represent inputs to knowledge that do not in and of themselves carry human-understandable meaning. Where those isolated elements come together into human comprehension, we tend to apply the word information; where information coalesces into a comprehensible narrative, we refer to knowledge. So even data science requires human intervention, most commonly by applying a narrative, in order to make the leap from data to applicable knowledge. Narrative, however, was viewed with suspicion by computer science researchers, who characterised narrative as ‘fake,’ ‘mostly not false, but they are all made up’ or, from a very different perspective, as a sort of ‘metadata.’ These researchers also expressed concern that peers might ‘pick the data to suit their story.’ Humanists and social scientists had a more nuanced understanding of the relationship between sources and scientific narratives, and of the balance between subjectivity and objectivity in their work. This emerged as one of the most interesting avenues for further work discovered by the KPLEX project, with particular resonance in an era of so-called ‘fake’ news and ‘fake’ science.

I. THE DARK SIDE OF CONTEXT: DARK LINKING AND DE-ANONYMIZATION

Clearly, the fact that big data is used at a distance from the context of its creation is a real and significant concern. But the loss of context is only half of the worry, as sometimes information that is supposed to have been removed is, in fact, indirectly visible. This threat of the preservation of unwanted context can be understood in terms of what is called ‘dark data’ or ‘dark linking.’ Given that we cannot necessarily know all of what data are available, we also cannot know where or how the identifying characteristics in even anonymised data can be re-established via proxies or triangulations. Digital discoverability therefore magnifies a dark side of data access that archivists were traditionally used to mediating as gatekeepers of material that is vulnerable to misuse. Although many of the computer science KPLEX interviewees were quite eloquent in their explanation of how we need to know ‘the purposes of the data. And the research. And the source. And the curation of it,’ they also knew of the potential for and cases of misuse, where data acquired within one project, or for a specific purpose, might be used or exploited by others for other purposes, or that consent given for reuse of personal data may be inferred or taken for granted, rather than explicitly sought.

Further research is required to deepen understanding of practitioners’ fears about the possibilities of data linking – and to examine the validity of these concerns within the uncertain future of the use of big data.

J. ORGANISATIONAL AND PROFESSIONAL PRACTICES

The need for organisational adaptation to big data methods was featured across the tasks and contexts investigated in the KPLEX project. To foster such changes (in archives, but also in universities and companies) we will need intermediaries, or perhaps translators, to ease the changes and ensure widespread benefit. Many interviewees and survey respondents pointed toward the need for such a skill set, and those who had experience of working with such people recognised their value: ‘[the State Archives] have someone who was an engineer at the beginning, but who is really capable to understand all the ways that archives work and the concept of metadata and [working with them] helps us to answer some technical problems.’ Such changes may be found already in the push toward the development of a large cohort of data scientists, but often the nature of and vision for such positions is quite limited, focussing more on data preservation and management than on facilitating new forms of exchange. In general, the competencies acquired in interdisciplinary research groups have not informed data science training programs, which could benefit greatly from the reflective elements of social science or humanistic knowledge. Fostering both more structured data management and a stronger convergence between traditional approaches and their datafied equivalents, present pressing needs in all of the sectors and contexts at which KPLEX looked.

K. BIG DATA RESEARCH AND SOCIAL CONFIDENCE

The fact that researchers, companies and memory institutions all struggle with big data platforms, research, and its results, points toward an even more widespread hesitation among citizens at large. Big data can be a powerful tool for knowledge creation, but power builds in-groups versus out-groups, and fosters a perception that pre-digital knowledge creation enables greater individual agency. Such a lack of faith has only been increased by the news coverage of corporate abuse or lack of care concerning personal data, and the threats to liberty, privacy, identity and democracy that have ensued. The scale of what big data platforms capture about us is astonishing, and what these same platforms may be denying us, in terms of access to the richness of our cultures, the diversity of our societies, and the range of perspectives we need to make informed decisions, perhaps even more so. Data literacy is not made attractive, neither by all-too-simple interfaces and platforms that offer results without complexity, nor by the lack of agency many feel in the face of big data. This is not a question of learning to code, so much as one of feeling empowered and included in the development of the digital society, and feeling that the digital world stands as an enhancement to our rich sensory and information lives, rather than in opposition to it. In fact, it should not be a question of citizens 'learning' much at all: technology should serve the aims of society, rather than creating a new category of invisible labour. Instead, big data platforms and products must make their biases and limitations clear, and assist the user not only to reach an end, but to grasp the means leading there.

Alongside these fears and hesitations about big data methods, the digital itself is an object of mistrust, not thought to be there for the long term. Some of this may be related to the intangibility of the digital, which resists basic human instincts about permanence, but underlying this is also a more accurate and oft-belied recognition of the fact that the digital transition is not a process with an endpoint. Even for historical documents, there is no such thing as a one-time investment in digitisation, and the need to continue to migrate formats and improve platforms implies that this transition may peak, but never be complete.

IV. RECOMMENDATIONS

The twelve areas for further research described above all point toward a forking in the road of big data research. The signs are already clear that to continue on as we have done, with the technological possibility to build leaping ahead of the human capacity to use to their own benefit, is incurring unsustainable costs. Incremental shifts have been suggested, such as industry focussing on privacy protecting technologies, or indeed even the funding of projects like KPLEX. But such shifts will do little in the long run to truly realign the trajectory of big data research, and the next set of changes that the algorithmic revolution in artificial intelligence is already bringing will only exacerbate the difficulties, and increase the inherent unconscious biases. In the place of incremental change, the KPLEX results point toward four possible areas of quite radical intervention into how knowledge creation pathways might be re-construed for the next generation of big data and society. Such measures will take courage to pursue, and their likelihood of upsetting extant hierarchies and power relationships will meet with resistance. The opportunity they could bring to re-establish the foundational assumptions of big data research could, however, be transformational, for technological as well as social development.

As with the development of technology itself, many of the things we need to facilitate this transformation are means and not ends: overcoming our unconscious biases, imagining the origins and destinations of our data, seeing the people behind the code. Such processes require more than a single initiative or intervention. But even a radical process must have a point of departure, and these are the four suggested by the results of the KPLEX project.

1. Enhancing regulation of big data research

KPLEX is not the only research project to come to the conclusion that software development and deployment, like driving a car or selling pharmaceuticals, can cause enough harm to certain users that regulatory responses should be considered. The European General Data Protection Regulation (GDPR) is a start, but perhaps only that. Only through regulation can access to public goods like data be secured, and only through regulation can potential harm be avoided. This should apply not only to breaches in the privacy of individuals, but also to gaps in access to the building blocks of strong, positive identities. Far greater transparency is required to be able to trace where big data comes from, how it is being used or manipulated, and who is responsible for and/or profiting from it. It is far too easy to lose sight of the human beings behind the data, and this potential for harm is as real as that within the context of such other regulated industries as air transport or power generation.

2. Rethinking the disciplines that contribute to big data research

The datafication of research requires us to rethink how we create the problem-solving toolkits with which we equip people at every level of society. This is not to say, however, that ICT skills should be taught at a younger age or made mandatory. Life in the 21st century requires a range of problem solving approaches. While it is important, perhaps, that experts in culture learn to code, it is equally important, if not more so, that engineers and computer scientists develop their ethical senses, their narrative imaginations, their cultural competencies, and their sensitivities to the communications skills they themselves deploy and respond to in others. This is not about ICT skills per se, but about critical thinking, as ICT skills will not support enhanced sensitivity to the limitations and reductions inherent in datafication processes, acknowledge the circumstances of data collection, or appreciate the potential biases of those creating knowledge. Most importantly, the ability to create knowledge via multiple input channels, as in interdisciplinary research, should become a core skill for every discipline and profession. Understanding the limits and potential of ICT must become a foundational skill, regardless of the context in which it will be applied, and in particular those who are training to become data scientists will need a very broad foundation to allow them to be maximally effective. Universities, active researchers, and professional societies would all have a role to play in this transformation, as would the publishers and funders that manage the incentive systems at the top of the research ecosystem.

3. Reversing knowledge hierarchies within big data research to disrupt biases and fixed mindsets

The mainstreaming of social sciences and humanities across the Horizon 2020 funding programme has the potential to become a key differentiator and source of innovation for Europe. This potential will not be reached, however, unless power imbalances and mismatches between large and small disciplines in projects are actively addressed. The European Commission and other research funders can assist in supporting such changes through the development of instruments to support fundamental integration, eg. through training, toolkits, restrictions on mono-disciplinarity in deliverables, recognition for ‘soft’ researcher skills as equal (if not superior) to hard innovation targets like patents, and other mechanisms by which to lower resistance, change expectations around research results and promote integrative knowledge creation. If funders signal the importance of this shift in mindset, the best researchers will be able to follow this lead. The system does require disruption to achieve this, however, but this can be managed by incentivising collaborations in which the minority perspectives must lead the work, and therefore cannot be marginalised. In addition, humanities and ICT communities need to be encouraged and facilitated to come together around some of the key concepts where the arts and humanities may be able to provide not only a social consciousness, but new perspectives on issues that an engineering approach might tend to excise from the process of knowledge creation: uncertainty, ambiguity, multiple perspectives, rich and contradictory narratives. Revisiting such limiting factors from first principles may well be the instigator for not only social and cultural innovation, but technological as well.

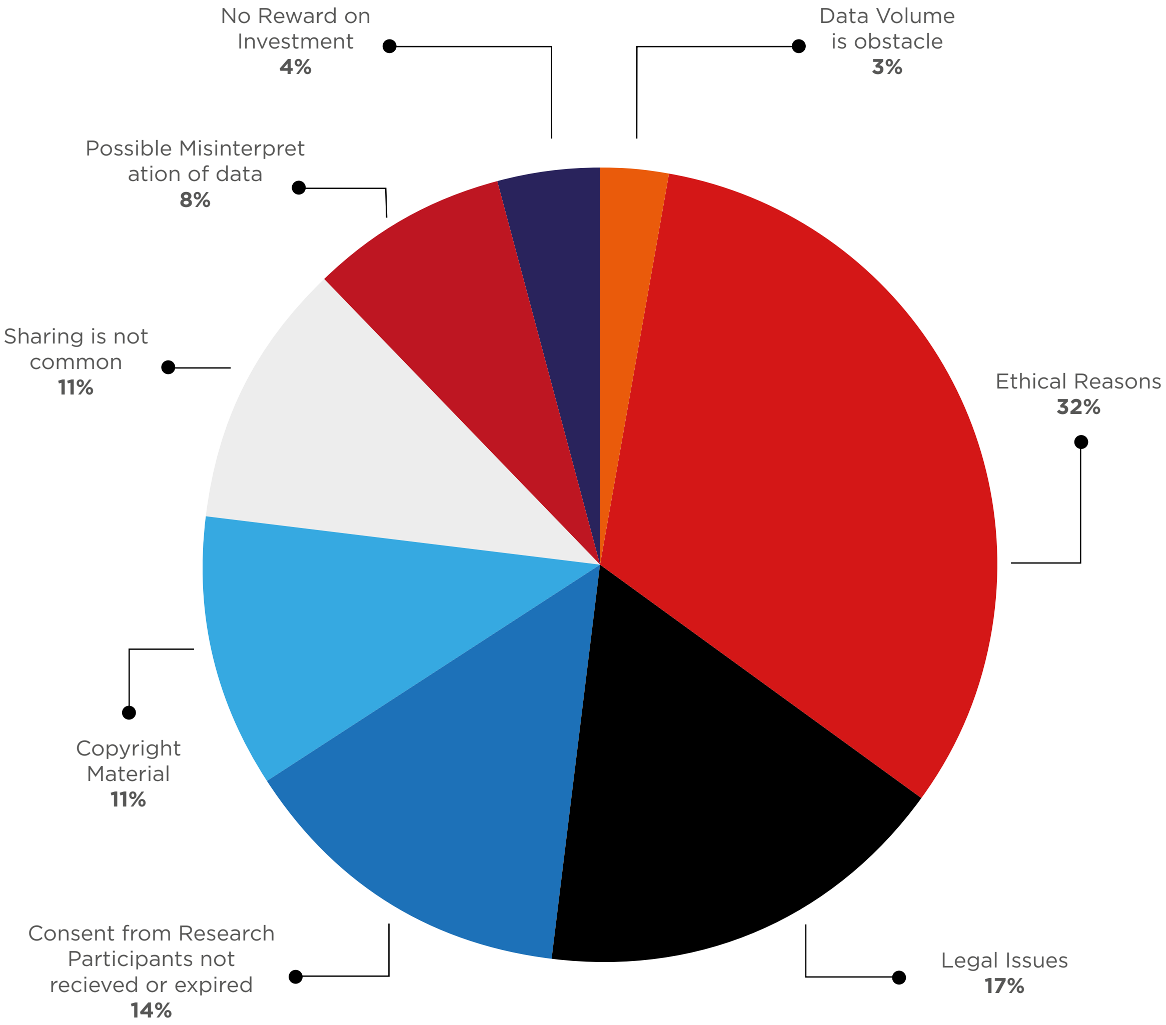
4. Ensuring contextualised data sharing for big data research, keeping context as minimal as necessary, but as rich as possible

Access to data is important, as is access to the contextual information that allows data to be used sensitively and reliably. This is a project on which the European Commission is already working, but if the European Open Science Cloud (EOSC) is to meet its potential, then far more preparatory work will be needed than just to establish a governance and technical framework. More sensitive instruments need to be developed for the capture and preservation of provenance and context, for example via a 'passport' style approach to research data, whereby successive iterations or applications of data, including rich source information, original formats and records of transformations and applications, can be captured and made available. Such a system should record not just a baseline standard metadata set, but also informal, contextual information: where has this data been used? Is there any further information available about it? Are there open questions about its origin?

This will be important as well in the harnessing of the long tail of research. The development of the EOSC, for example, will lose greatly in its richness if the current conceptions of data continue to be so divergent, with some disciplines using the word to mean a wide variety of things, and many others seeing it as an irrelevant term for their work. If data is to be understood as a fundamental, basic building block of interdisciplinary enquiry, much work will need to be done to develop greater consensus around this term among disciplines currently very far apart in this matter. All disciplines must be encouraged to see the richness of the data layers in their work, and all researchers must be incentivised to share, for preparing data for reuse is laborious, and does not have an immediate return for the scientists undertaking such work. An expanded role for research libraries should be considered in this respect, as well as for workflows that harness, rather than replicate or come into conflict with, existing research processes. Only with such a multifaceted approach can the many reasons given for not sharing data (see image) be countered.

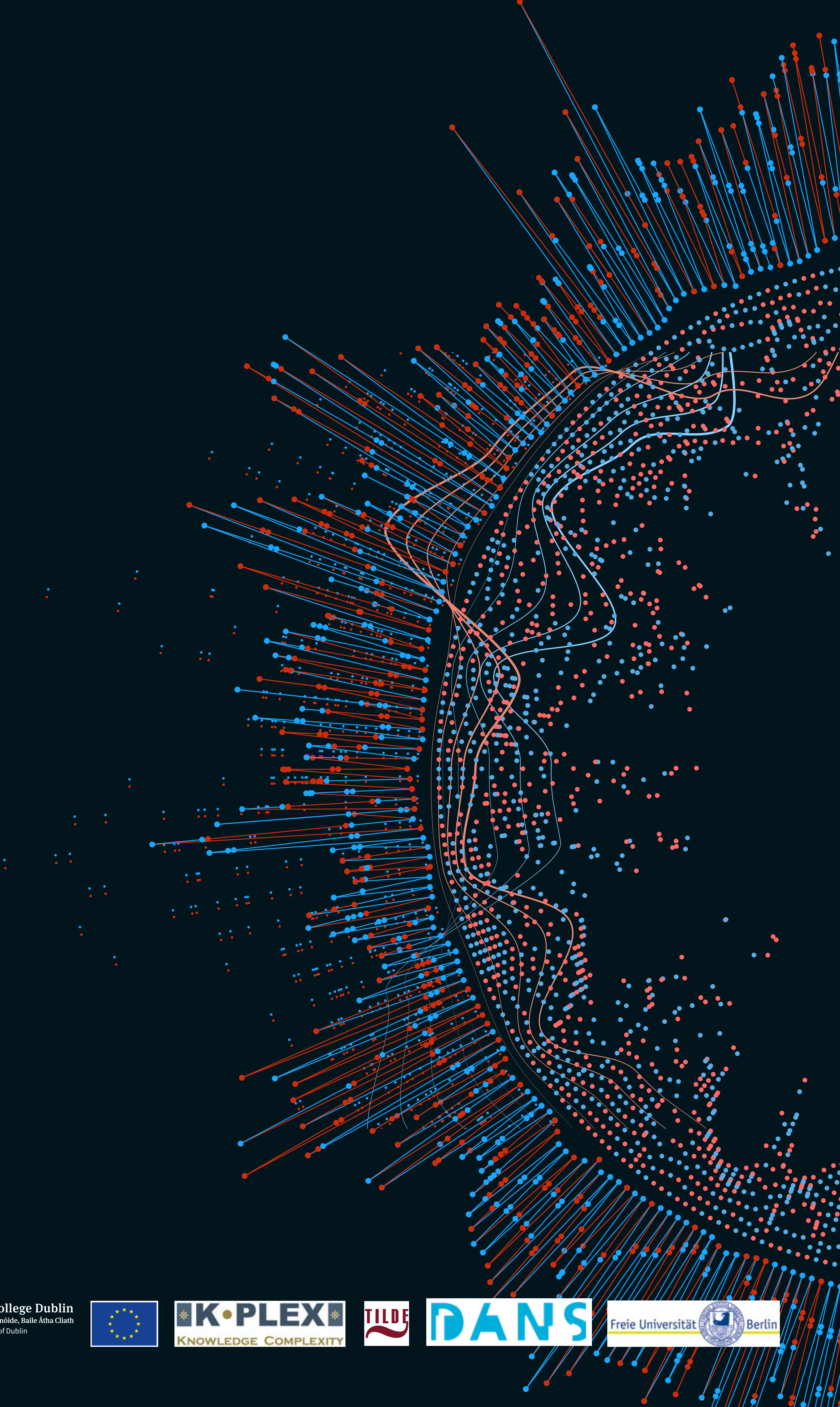
Finally, the vision of the EOSC, that research results be available to academic, public and industrial users, should not be a one-way street. Industry too should share their data, as should organisations in key fields where data may be lacking, such as professional organisations and tourism boards.

Reasons for not sharing data:



V. IMPACT OF THE KPLEX PROJECT

As a 'sister' project intended to undertake research linked to other Horizon 2020-funded big data research areas, KPLEX itself did not have either the time or resources to fully develop the many potential interventions its exploratory research suggested could be implemented to reduce bias and increase richness in big data research. We can hope, however, that future research and policy development will encourage the big data research community as a whole to take these opportunities to rethink fundamental assumptions and foster a more symbiotic relationship between technological and social progress. This will be a necessary development for Europe if the recognised risks of big data research are to be countered at the macro level.



Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

K•PLEX
Knowledge Complexity

TILDE

DANS

Freie Universität Berlin