Deliverable Title: D1.1 Final report on the exploitation, translation and reuse potential for project results

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

▶ **To cite this version:**

Jennifer Edmond, Michelle Doran, Nicola Horsley, Elisabeth Huber, Rihards Kalnins, et al.. Deliverable Title: D1.1 Final report on the exploitation, translation and reuse potential for project results. [Research Report] Trinity College Dublin. 2018. hal-01842365

HAL Id: hal-01842365
https://hal.science/hal-01842365

Submitted on 18 Jul 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Deliverable Title: D1.1 Final report on the exploitation, translation and reuse potential for project results
Deliverable Date: 30th March 2018

Project Acronym :	KPLEX
Project Title:	Knowledge Complexity
Funding Scheme:	H2020-ICT-2016-1
Grant Agreement number:	732340
Project Coordinator:	Dr. Jennifer Edmond (edmondj@tcd.ie)
Project Management Contact:	Michelle Doran (doranm1@tcd.ie)
Project Start Date:	01 January 2017
Project End Date:	31 March 2018
WP No.:	1
WP Leader:	Jennifer Edmond
Authors and Contributors:	Jennifer Edmond (edmondj@tcd.ie) Nicola Horsley Elisabeth Huber Rihards Kalnins Joerg Lehman Georgina Nugent-Folan Mike Priddy

	Thomas Stodulka
Dissemination Level:	PU
Nature of Deliverable:	R = report
Abstract:	This report provides an overview of the results of the KPLEX project, focussing in particular on the insights, conclusions and recommendations developed by the project which will be most relevant for future reuse by other research groups and sectors. It outlines the key gaps the project discovered in big data research, and proposes a series of concrete measured that should be taken to address them.
Revision History:	Version 1.0 uploaded to the EC portal, March 2018 Version 2.0 was revised to include additional explanatory text on the project implementation strategy (page 7-8) and uploaded to the EC portal in July, 2018.

KPLEX Project, Draft Exploitation Plan

Version 2.0

Table of Contents

I. Executive Summary

This document fulfills the Deliverable 1.1, “Final report on the exploitation, translation and reuse potential for project results” of the Knowledge Complexity (or KPLEX) project, funded by the European Commission under Grant Agreement number 732340. The report lays out how the results of the KPLEX project may be reused by a variety of actors within the big data research ecosystem.

The report presents some basic reflections on the project’s method and composition, as well as three different views of the project’s results. First, it describes the conclusions of each task-based work package in the project. Second, it presents a series of high-level issues documented by the project about the complexity of human activity and its resistence to accurate representation within big data systems. Finally, the report gives four concrete recommendations for how big data research can be enriched and strengthened in the future: enhancing regulation of big data research, rethinking the disciplines that contribute to big data research, reorganising knowledge hierarchies within big data research, and ensuring contextualised data sharing for big data research.

II. Introduction

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 by harnessing methods and challenges situated in a research community that has been relatively resistant to big data, namely the humanities. The project’s fundamental supposition, which has been borne out by the experiences of delivering on its objectives, has been that there are epistemic and practical reasons why humanities research resists datafication, a process generally understood as the substitution of original state research objects 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 such areas of epistemic divergence, and, from this unique perspective, propose where further work could and should be done.

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 also harnessing 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.

III. Audiences for the KPLEX results

Given the broad aims and objectives of the project as defined in the introduction to this document, 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, including an industry-based one. The techniques by which we have both differentiated and aligned our epistemic 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. Each of these areas will be able to draw from the project 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 imagination are becoming ever more widely recognised in ICT development and regulation, KPLEX's results should be a welcome source of fresh thinking, encouraging 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 operating. People know enough about big data research to be concerned by it, but the interdisciplinarity 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. 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 the pre-digital mission on the other.

All of these audiences have been engaged, to varying degrees, in the course of the KPLEX project's lifecycle, and the project's final Deliverables will provide a rich resource base according to which this resonance can continue long after the project has closed.

IV. The KPLEX Consortium and Approach

The KPLEX project team was recruited in such a way as to be both 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 disciplinary community, with different foundational expectations of and from the knowledge creation process: the Coordinator (Trinity College Dublin, TCD) is based in a traditional, historic university and, more specifically in their arts and humanities research institute, but works in the area of digital humanities, and therefore would have been experienced in interdisciplinary approaches. Her research team of two had a similar profile, both qualified in traditional literary/historical approaches, but with some exposure to collaborative work through an interest in and engagement with digital humanities. The digital humanities provided a good approach from which to instigate this particular project, given its fundamental interdisciplinarity, technological awareness, and tendency toward fluid disciplinary identities.

The Freie Universitaet Berlin (FUB) team also enjoyed the benefits of being based in a large and broad-based university. This team's disciplinary orientation was quite different from TCD, however, representing a fundamental grounding in anthropology and ethnology, which in many

places would be considered a social science, rather than a humanistic discipline. The FUB team was somewhat more diverse, however, including two researchers self-identifying as anthropologists (including the PI), and a third (originally trained in literary studies) brought in specifically for his interest in the digital aspects of the work. Anthropology was an essential addition to the project, as the need to observe human behaviours and emotions was a key enabler for the project.

DANS is a very different sort of institution from either the FUB or TCD. As a smaller organisation, more focussed, and with a primary identity as a research data archive, DANS brought a very practical approach to the project, but also an excellent understanding of the institutional forces leading to barriers and restrictions we were likely to find. The team was led by a computer scientist, and augmented with a sociologist. Their experience of delivering data and data-related services gave us an excellent grounding for understanding how cultural big data is managed, where fissures are appearing in current practises, and how these may be different from those in other fields.

TILDE was perhaps the most different partner of them all. As an SME, rather than a public institution, they injected into the project a perspective that was much more focussed on their sector and their needs as a private sector organisation, a key potential blind spot for a research project led largely by basic researchers. They also gave us access to insights from across their networks, looking into the imperatives of commercially-funded research and products needing to reach a consumer audience.

At the outset of the project, all of the partners had had experience of collaborative, interdisciplinary research, and all had known the coordinator, but none had worked directly together before. The need to build a team out of these diverse perspectives was a great challenge, which was addressed by KPLEX with a higher than usual density of four face-to-face and eight virtual meetings over the course of 15 months.

The project faced two constraints that shaped its overall approach and methodology. First, it was such an uncommonly short project that it needed to be managed very efficiently in order to achieve meaningful results. Second, the partners brought very diverse expertise and frames of reference to the project. This was a strength, but also meant that we needed to be mindful of the fact that results would need to be transferred both according to disciplinary or community of practice lines (not least in order to ensure that the early career researchers were able to translate their work into usable academic capital for their continuing careers) as well to a broader, cross-disciplinary and indeed cross-sectorial audience.

The best response to these constraints was determined to be to keep the focus and organisation of work spread across the four key questions described below and across the four institutional teams described above. The four teams did not operate as silos, however: in addition to the large number of meetings and exchanges built in to the work plan (described above), we also staged work in such a way to facilitate integration. For example, the initial draft WP 2 white paper on varying definitions of the word data and other key terms (which now forms

the core of the Deliverable 2.1) was available already by the time of the team's second substantive face-to-face meeting in month 6. This was a remarkable achievement given that the first three months of the project were given over to recruitment of the research team. To give a second major example, the research team also built a set of overlapping codebooks to allow for an eventual comparative study of themes and topics in the various surveys and interviews conducted.

Early on in the project, we did develop templates for researcher diaries that would explicitly document our potentially divergent knowledge creation processes. This proved challenging, however, as such instruments were too far removed from many of the researchers' regular practices, and were difficult to integrate and maintain meaningfully. We therefore fell back on to the socratic method of achieving integration, fostering formal and informal conversive methods for developing a shared vocabulary, a shared set of themes and a shared set of conclusions and recommendations. We set ourselves challenges beyond the capacity of any one partner to deliver, and trusted each other to deliver to the best of each contributor's own expertise. This did not mean that the integration was seamless, however. For example, even we often described phenomena with different words: what WP2 called "narrative," WP3 identified as "context;" what WP2 called "a loss of provenance," WP 4 called "a loss of information." The richness of our sharing mechanisms made it possible for us to recognise and resolve these discursive variations, however, and redress their potential to hide overlaps between the insights and focus areas of the teams.

Such ongoing negotiation of key terminology and hierarchies was a continuous theme running through the KPLEX project, and made the project challenging but deeply enriching. We were greatly fortunate, however, to have a group assembled who also projected personal qualities of confidence in their own work, a desire to see their work from other perspectives, and an impatience with the power dynamics so common in research collaborations. The partners came to the project from a default position of respecting each other, which, over time, grew to greatly enjoying the always stimulating, but not always easy, conversations KPLEX fostered. This theme will return in our recommendations, but the success of KPLEX is a strong reminder that both intellectual and personal qualities play in to the best collaborations.

V. Summary project findings by theme and task

The KPLEX project was originally structured around four key questions. The method by which the project sought to answer these questions involved a large and multiperspectival information gathering exercise, covering four distinct surveys, tens of hours of interviews, and a data mining exercise. The groups from which participants were drawn included a range of perspectives from around the practice of big data research, including cultural heritage practitioners, researchers from humanities, science and engineering, and technology developers. The founding questions and a brief summary of the project's progress beyond the state of the art in each of these initially defined areas is given below. Please note that a much fuller and detailed discussion of the work on each project theme can be found in the related workpackage deliverables.

Task 1: Redefining what data is and the terms we use to speak of it. *In specific, this task sought to develop an understanding of how computer scientists speak of and view the data they work with, to better understand where and how the dissociation of the data set from its origins in individualised processes of gathering information takes place. It was intended as a transferable piece of work that could inform the other strands of activity.*

The initial literature review for this task confirmed that even philosophers and sociologists of science and technology could not agree a stable definition of the term 'data.' Through the course of a series of interviews with computer scientists and a data mining exercise looking at published big data research, we were able to confirm and expand upon the conclusions that people define the word data differently, people interpret data differently, and that the term data is heavily overdetermined. The inconsistencies of definitions and variability of what data can be, how it can be spoken of, and what can or cannot be done with it, were striking, and the sheer scale and variance has significance. For example, it is possible to have:

> *Ambiguous data, Anonymous data, Bad data, Bilingual data, Chaotic data, Contradictory data, Computational data, Dark data, Hidden data, Implicit data, In-domain data, Inaccurate data, Inconclusive data, "Information" (rather confusingly, one interviewee referred to data that is "understood" to have meaning as "information"), Log data, Loose data, Machine processable data, Missing data, Noisy data, Outdated data, Personal data, Primary data, Real data, Repurposed data, Rich Data, Standardised data, Semi-structured data, Sensitive data, Stored data, True data, Uncertain data.*

Each of these terms has different implications for the state of data, where it comes from, what can be done with it, what has already been done to it, what its status, or indeed 'truth-value,' is as a representation of particular phenomena. Consistency of definitions and the semantics of the term data were more notable among researchers with the same disciplinary training (such as computational linguists), among researchers working on the same project (who would have established over time clearer communication practices regarding the working definitions specific to that project), or researchers on a small team who were familiar with each other, and had developed a more intuitive understanding of each other's research habits, methods, and preferences.

The work of this KPLEX task also developed new insight into the knowledge creation process as it is pursued in big data research. Terminological comorbidity and counter-narratives to the traditional 'data-information-knowledge-wisdom' (or DIKW) hierarchy were observed throughout, particularly between the terms data and information, with "data" being referred to as something that contains "information" (but also, perhaps, vice versa). Further still, the same term is used to refer to different kinds or states of data within one overarching research project (data stream, data cluster, original data, evolved data, evolving data), and as a synecdochal term covering both the whole and the part. The word was also often used to refer both to specific data/ datasets & more general data / datasets, and to refer to data from different phases of an investigation: The data source transitions from being pre-processed data, to processed data,

and then becomes “raw” data that, with further data processing and data analysis, will lead to output data. This presents as a problem that is surmountable on a small scale, or even at disciplinary level where commonsense, familiarity, and a shared disciplinary background renders the interlocutors fluent in and familiar with each other’s terminologies and research methods. But our traditional reliance on community ties to overcome the flaws in both our data and the terminology we use to speak of it do not translate well to larger scale interdisciplinary endeavours, to environments where the backgrounds or motivations of researchers/ participants are not necessarily known or trusted or to environments where either the foundations of the research objects (such as is found in big data) or those of the algorithmic processing results (such as found in many AI applications) are not superficially legible to a human researcher.

Many of the interviewees, particularly those with experience in interdisciplinary research or mediating between disciplines, such as computational linguists, expressed concern 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 abilities of alternately engineers or humanities researchers to fully comprehend what they were working with. More specific terminology, more detailed provenance records, greater transparency regarding the transformations applied to data, and increased accountability on the part of those responsible for the curation of data will be required help counteract this distrust and dissuade insular or obfuscatory behaviour.

Researcher accountability was highlighted not only as a key factor in assessing the validity or reliability of data, but also as a key concern among researchers. The need to resist cherry picking data to adhere to pre-established research hypotheses or desired narratives, was also repeatedly stressed. The provision of adequate training and continuing professional development regarding best practice with regards to data curation, processing, and analysis are additional areas where KPLEX has exposed the need for improvement. Regarding complexity in data, context and disambiguation emerged as the two key processes that, when combined, serve to effectively tackle ambiguity or uncertainty in data: that said, as the work of other tasks will show, maintaining context and clarity in an environment optimised for processing is seldom without compromise, and many of these compromises carry unseen downstream consequences.

Task 2: ‘Hidden’ data. *This task had as its focus the need to come to grips through data with the analogue world of the individual, where incompatible information streams, some analogue and some digital, are seamlessly combined to make knowledge. In particular, it looked at how the socio-technical nature of knowledge environments impacts upon the development of identities and in particular at historical research processes. It probed the common assumption that all data of value is, or will soon enough be, available in digital format, a presumption that is widespread and which professionals who measure their collections in metres and kilometres strongly contest. In particular it looked at how such tacit assumptions about access to data and information put the development of a holistic and balanced understanding of the past at risk.*

The most instructive lens through which to view the challenge of the hiding or obscuring of analogue data is through the changing practices and observations of, and the challenges faced by, memory institutions (the original big data organisations) such as museums, libraries and archives. Collection digitisation and data aggregation systems have promised to expand access to cultural heritage collections, or at least to the metadata describing them, and increase the possibilities of archival research. The difficulties of this seemingly straightforward task of bringing together the data of disparate and distinct institutions have been well documented, however. That the move from analogue to digital has been embraced, at least at some level, across the archival sector, is therefore no trivial development, meaning that even where the organisation of material has not really changed in living memory, the digital revolution is mandating that institutions revisit their fundamental practices. The challenge is greater than just the production of digital catalogues and surrogates, however, necessitating work like back cataloguing or updating metadata, work that had been low priority until the digital turn created risks that improperly described collections could slip into irrelevance or obscurity.

Shaking up institutional practice from the outside, for example through infrastructure projects, sometimes achieves significant change in a relatively short time. External influences were cited by many KPLEX interviewees as the catalyst for adopting greater standardisation of item descriptions but this process carried the risk that the “very rich” original descriptions of holdings would disappear from the historical record in favour of those that were compatible with other archives. Archivists described how they strove to “save” this knowledge by mapping original metadata to the standardised schema, a labour-intensive process. Dealing with items that had not necessarily been collected as part of a conscious, archival vision was a particular challenge. Preserving the context of a resource through metadata was therefore revealed as a core concern among archivists, in particular as their understanding of what had influenced the representation of items at different points in time (often pre-dating the archival institution itself) was critical to interpreting how it might be used in relation to other research resources.

Metadata was seen as one of the vehicles for conveying context to the user but researchers’ experience of exploring resources was also seen to be significant. There was broad acknowledgement that the changing use of collections disrupted some of the fundamental tenets of archival and cultural heritage preservation practice. For instance, the traditional hierarchical structure was reported to be losing significance as it was undermined by Google-style keyword searching. There was some resistance to fully aligning with the new paradigm, however, and rivalling Google was an ambition of many participants to counter the danger of sources being hidden by an unseen algorithm: “to be stronger than Google”.

It was felt that the influence of Google went beyond the user-friendly interface of the search box as the culture of search engine use was like a self-driving car, but one where the destination would not be a fixed point, and you couldn’t ‘control on what basis the machine gives the result.’ Looking to the near future, participants envisioned research methods becoming further removed from the researcher’s hand as automated tools, machine learning and AI would play an increased role. There was also a concomitant fear that increasing data linking could identify previously anonymous data subjects, revealing data that had been ‘hidden’ for ethical reasons.

Furthermore, additional risks were identified that in subordinating knowledge complexity to corporate interests, there could also be an overall reduction in the scope of research. It was generally agreed that intervention would be necessary to maintain appropriate standards of rigour in incipient generations of researchers to allow them to maintain control over their own research and not fall victim to the blind spots created by a curated "cultural Facebook." Practitioners therefore supported efforts to open up the full range of cultural heritage knowledge at every stage of the research process. In order to achieve this goal, there would need to be a greater appreciation amongst all those working with heritage institutions of the potential for data to become hidden due to both the unintended consequences of sharing and the desire to share being blocked by a lack of appropriate resources and tools.

Task 3: Knowledge organisation and epistemics. *This task looked at scholarly research data and in particular humanities research data so as to understand how complex data sets are shaped by the interpretive mind that creates the organisation framework, even (and especially) in a landscape where standards for such organisation exist, but may be applied quixotically or indeed ignored.*

To address this question, KPLEX researchers looked at the practices and attitudes of researchers focussed on the topic of emotion, including a broad range of scientific, humanistic and social science disciplines, epistemologies and methodologies, including how they are managed and organised. The findings show that emotion researchers are well aware of the fact that datafication is connected to loss of information. Data structures and data collection are all highly dependent on research epistemologies, research questions, and the methodologies and technologies used in order to create scientific knowledge, and therefore highly malleable.

The multitude of disciplinary, methodological and theoretical approaches used in emotion research engenders a heterogeneity of data structures and formats. Bias results from the limitations in coverage of the chosen approach, in particular as data are collected at different points in time during the research process, and are often dependent on the researcher, since the researcher acts as an instrument in the data collection process. Moreover, scientific researchers often equate datafication with quantification, which is not necessarily the case since ethnographic field notes or textualization of systematic observations can also be described as a form of datafication.

Because research on emotions is not confined to a single discipline, interdisciplinary research projects are common. Reflecting on their experiences, the research participants underlined the need for a comprehensive theory as a basis for data integration. According to this paradigm, data structures should be conceptualized according to the aims of the research questions, but not necessarily from the point of view of reuse. As a prerequisite for data reuse, participants therefore identified the knowledge on the context in which data were collected, such as research questions and methodologies, as a key requirement, as well as cultural or language peculiarities of the research setting. It might be expected that use of digital methods might provide a lingua franca for researchers from different backgrounds approaching similar topics, but to the contrary, even within academia, fluency with ICT-based terminology could not be taken for

granted. In fact, within the humanities and some social sciences, a discursive resistance to datafication was noted, since the terms “research material”, “narratives”, “stories” or “sources” were used instead of “data”, especially where the data used was not bespoke data created as a part of the research process (eg. literary texts or historical records rather than researcher-designed surveys or interviews). Texts and transcripts were not considered to be data, and standard procedures like quality checks for data formats were widely unknown.

Regarding big data as a knowledge creation model, researchers expressed their uneasiness as well as their uncertainty concerning the definition of big data and the consequences big data research might have for scientific epistemologies and methodologies. Big data driven research does not emerge from established research processes, i.e. having a research question, as well as structuring, collecting, and analysing data before writing up the results. The fact that big data can currently only be analysed in an exploratory mode was criticized by scientific researchers who are used to the implementation of carefully guided, methodologically and epistemologically comprehensible research practices. On the other hand, this line of questioning also exposed particular strengths of the humanities, the ethnographic and qualitative social sciences that should be contributing to big data analyses, such as sensitivity to the contexts in which data were collected, dealing with the non-representativeness of big data, or the potential societal repercussions that could emerge from the classifications contained in large datasets.

The multiplicity of epistemic approaches and the tendency of academic disciplines to demarcate themselves from each other as part of their struggles for legitimacy and recognition led to asymmetric power structures and exclusive approaches within academia. Research on emotions shows in an exemplary way that these structural factors result in a marginalisation of complexity. Interdisciplinary research projects can provide a response to this, in particular if they are constructed so as to start with an encompassing theory, implement methodological reflections across disciplines, and conceptualize data depositing strategies from the point of data reuse.

Task 4: Culture and systemic limitations. *This final task looked at how big data approaches and systems flatten, misrepresent and potentially threaten the viability of aspects of culture within cultural data systems. From a technological point of view, this task focused on restrictions to the cultural diversity of big data systems and on how they make their underlying framework assumptions (such as database relations or ontologies, search or personalisation algorithms, etc.) and the limitations of their reliability clear.*

The language technology (LT) industry was chosen as a test case for examining issues surrounding big data and its availability, as well as the impact of data on technology, infrastructure, and employment. LT solutions like machine translation (MT) are developed with culturally-specific language data as input material, therefore data quality issues, such as errors, noise, and inconsistencies in coverage, can have a crucial impact on the quality of services. These problems are exacerbated by data scarcity and inequality, particularly for smaller languages and overlooked domains.

To explore the impact of language data availability on translation technologies, KPLEX researchers collected input from the LT industry on their use of LT solutions and related data issues in two comprehensive surveys, covering many contributors to the language technology community. These surveys found that MT systems are widely used in generic online translation services for the world's largest languages, but more than half of users said they were only partially satisfied with LT support for their native language, and 25% were not satisfied at all. The surveys also found that half of language resource (LR) specialists felt that the overall volume of LRs was insufficient to meet their needs. The most frequently encountered issues with LRs were data availability, openness of data, and Intellectual Property Rights (IPR) issues.

A lack of data processing tools for all EU languages, as well as the restricted availability or openness of data, were also identified as problems. The majority of LR specialists (80%) had encountered a lack of natural language tools (i.e., text processing tools, speech processing tools, and semantic analysis tools) to process data, crucial building blocks for developing LT solutions. Moreover, the vast majority of LR specialists (79%) had encountered LRs that they could not use for reasons of data availability, openness of data, or IPR issues. A majority of LR specialists (51%) had also had at some point to forego or turn down a project, research study, or other opportunity due to LR issues, most frequently on account of the availability or non-existence of LR. Survey respondents asserted that they considered the following elements "essential" in helping to overcoming language resource issues: more freely available public data, easier IPR clearance processes, and better language coverage. Helping to make more data, particularly public data, available for LT developers in all EU languages, as well as easing IPR clearance procedures, would therefore be a crucial step in providing better LT solutions for users.

Though survey respondents acknowledged that language data was impacting their business very heavily, and would continue to do so in the future, respondents also admitted that language data management processes at organizations are mostly loosely structured or ad hoc. Moreover, more than half of organizations did not employee data engineers, data specialists, or data managers on staff. This lack of capacity within the organisations working in LT added a further layer of concern that LT for languages other than the largest ones would continue to be a poor shadow of the possible, and that the cultural implications of this invisible cultural inequality would continue for some time to come without significant proactiveinvestment.

VI. Integrated findings and areas for future work

The previous section of this report demonstrates the quality of the insights generated by the individual reaserch themes pursued by the researchers of the KPLEX project. The most compelling outcomes of the project stand at the intersection of the perspectives and themes pursued by these individual work packages and tasks, however. These points, which cover a wide range if issues around the complexity of the phenomena represented in data and the potential biases inherent in the methods most commonly used to interrogate them, are described below. The resonances between and across the perspectives mined by the project

illuminate those 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 some 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 fulfillment of the technical to render complex phenomena in a binary system of 1s and 0s feeds into a very human attraction to answers that may be simple, straightforward, confident, and possibly even false, or at least misleading. Big data research tends to portray complexity as a negative, rather than a positive, and commits to the marginalisation, removal, structuring or 'cleaning' of complexity out of data. Human life, as with all organic processes, is, however, inherently complex and difficult to capture and represent effectively without reference to the context in which is a particular representation of it (such as a dataset) was conceived. The KPLEX project was able to observe the resistance among many information experts to simplification across its research themes. In particular, the emotion researchers who were interviewed were able to express elegantly those aspects of their research they would not be able to capture and quantify as data: phenomena 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 real challenge for representing many aspects of 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 encompasses the whole ambit of the data, 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. Metadata records in archives were viewed with some suspicion, in recognition of the fact that cataloguing records are 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 (EmotionML) in their descriptions and analyses.

Similarly, keyword searches 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 take this as one of their most important duties and and greatest challenges, to meet the optimal compromise in capturing context to support the appropriate use of their holdings. This is what they are trained to do, but it is 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.

Interestingly, the question of context represented one of the strongest points of divergence between researchers from different backgrounds, with computer scientists often characterising the cleaning of data to excise uncertain elements as a necessary and intrinsic part of the research process, while others saw this as data manipulation and antithetical to good research. Further research is required to investigate in more detail what if any price is being paid in these contentious processes.

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 is multidimensional situated in 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 reaching conclusions. 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. To go even further, the creation of data sources, such as archival descriptions or interview transcripts, is clearly perceived as 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 companies, 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 captured and held was 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 created dominant forms of information retrieval and knowledge production that, in spite of their inherent biases and limitations, were 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, and potentially without verification or indeed intended to mislead.

The digital record can suffer from the dominance of a single mode of access, but also from the impoverishment of 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 fieldnotes, 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, and of the strength of cultural ties to create resilient societies, then we should be very concerned about how the digital record, for all of its wonderful, 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 is 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 epistemic 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. The dependency of data on the context in which they were collected was often mentioned by our interviewees.

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, were 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 the 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 'you are interested in problems of data but this is not my kind of work.'

The variations among such differing applications 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 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 pidgins that meant the same thing. Language is not only about communication, 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 as much to do with evening out the engrained power dynamics and facilitating fundamental epistemic negotiations, such as an early negotiation of key terminology. Too many interdisciplinary projects proceed, perhaps through their entire life cycles, without ever developing the vehicular 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 were 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 areas of expertise 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 observe its objectivity or truth value

Big data was highly critiqued for its loss of 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 represent the inputs to knowledge that do not in and of themselves carry human-readable meaning. Where those isolated elements come together into human understanding, 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.' They also expressed concern that researchers might 'pick the data to suit their story.' Needless to say, 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.

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

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, 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 is available, we also cannot know where or how the identifying characteristics in even anonymised data can be re-established via proxies or trangulations. 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 or personal data may be inferred or taken for granted, rather than explicitly sought.

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

k. 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, but also a stronger convergence between traditional approaches and their datafied equivalents, present pressing needs in all of the sectors and contexts at which KPLEX looked.

I. 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 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 record about us as citizens 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, even more so. Data literacy is not made attractive, neither by all-too-simple interfaces and platforms that offer results stripped of complexity, nor by the lack of agency many feel in the fact 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, affective 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.

VII. 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 reconstrued for the next generation of big data. 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 reestablish 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 the building of 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 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, and their sensitivities to the communications skills they themselves deploy and see 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 in 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 epistemic 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 resistence, 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.

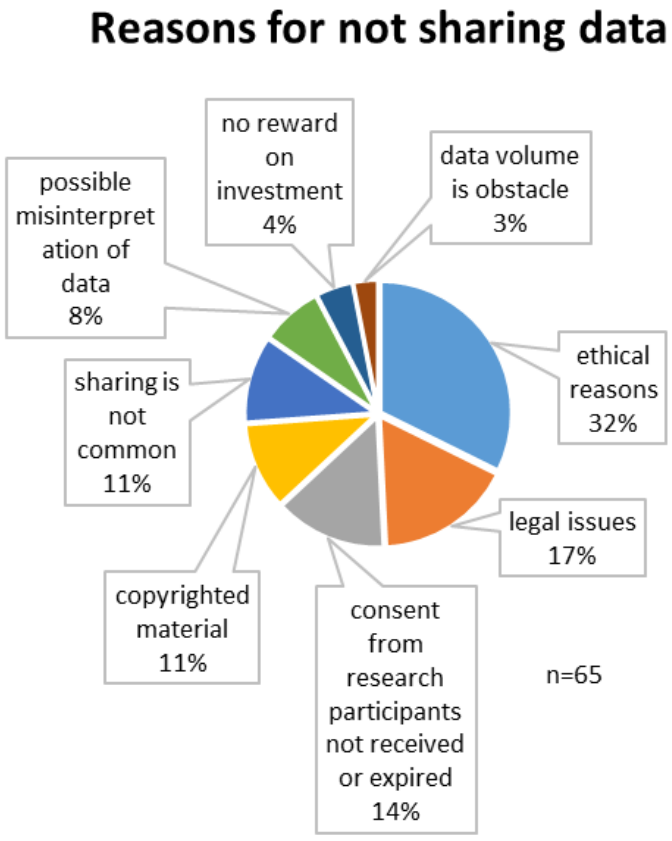
4. Ensuring contextalised data sharing for big data research, keeping context as minimal as necessary, but as rich as possible

Access to data is important, as is access to the contextual information that allows it 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 on 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.



VIII. Impact of the KPLEX project

As as 'sister' project intended to undertake research linked to other Horizon 2020 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. That said, we have been able to create a firm basis of empirical evidence and develop concrete recommendations for future development. Whether or not these will be taken up directly by the actors in the Big Data PPP is another question. The biases, both implicit and explicit, against such issues are strong in the community, and, after the close of KPLEX, such challenging perspectives as we have voiced will no longer be present in their discussions (to the extent that they have been entertained in the first place). We can hope, however, that future research and policy development might be able to encourage the big data research community as a whole to take the opportunities to rethink fundamental assumptions and foster a more symbiotic relationship between technological and social progress. If the recognised risks of big data research are to be countered at the macro level (rather than in closely bounded terms such as protection of individual privacy), this will be a necessary development for Europe.

Within its active grant phase, the KPLEX project shared its work with a wide variety of scientific audiences (including ICT researchers, humanists, science and technology studies, library science, data science and others) as well as the general public, through blog posts and popular media. The project engaged with cognate research projects and research users through both the Big Data Value Association and Hub-IT platform, and its partners are already looking to deepen and expand their work into areas such as AI and research infrastructure. In addition, it leaves the legacy of its fully contextualised, openly accessible research data. In spite of the project's short duration, this active dissemination, communication and exploitation programme will leave a legacy far beyond the scale of the project itself.