Deliverable Title: D2.1 Redefining what data is and the terms we use to speak of it.
Deliverable Date: 31st March 2018
Version: V2.0

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.:	2
WP Leader:	Dr Jennifer Edmond, TCD
Authors and Contributors (Name and email address):	Dr Jennifer Edmond (edmondj@tcd.ie) Dr Georgina Nugent-Folan (nugentfg@tcd.ie)
Dissemination Level:	PU
Nature of Deliverable:	R = report

Abstract:	How do we define data? Through a series of interviews with computer scientists and a text mining exercise, this report presents the investigations into the how data is defined and conceived of by computer scientists. By establishing a taxonomy of models underlying various conceptions of what data is, as well as a historical overview of when and how certain conceptions came to be dominant, and within what communities, this report moves towards a reconceptualisation of data that can accommodate knowledge complexity in big data environments.
Revision History:	Version 1.0 uploaded to the EC portal, March 2018 Version 2.0 was revised to include the KPLEX Project Glossary of Big Data Terms for Humanists (page 208-210) and uploaded to the EC portal in July, 2018.

Table of Contents:

1. WP Objectives

WP 2's objective was to deliver a fundamental comparative layer of information, which supported not only the research of the other WPs, but also enable project recommendations and underpin the discussion of D 1.1, the project final report on the exploitation, translation and reuse potential for project results. The objective was to synthesise findings into both a white paper for policy/general audiences and a journal article.

The activities of the WP2 were broken down into the following tasks.

T 2.1 Survey of the state of knowledge regarding the development of the concept of data (M 4-6)

Drawing on a wide range of disciplines (primarily history, philosophy, science and technology studies, but also engineering and computer science), this first task sought to establish a taxonomy of models underlying various conceptions of what data is, as well as a historical overview of when and how certain conceptions came to be dominant, and within what communities. This work allowed the project as a whole to base its further development within a contextualised frame.

T 2.2 Development of data mining exercise and interview questions (M 7-8)

At the conclusion of the first phase (and in conjunction with the consortium's midterm face-to-face meeting), the original objectives of WP2 as outlined in Annex 1 of the DOW were to develop both an online survey and a set of detailed questions to underpin one-on-one interviews with computer science researchers. As outlined in the interim project management report, which was approved by the project's European Project Manager, these objectives were amended to reflect WP2 aim which was to acquire rich rather than large responses to the accompanying WP interview. Therefore we amended our objective of conducting a large survey and in the place of a survey, WP2 conducted a data mining exercise across a corpus of computer science journals and proceedings from "big data" conferences so as to get a more informed picture of what inherent definitions of the word "data" are expressed in them, and how transformation processes like data cleaning or processing are viewed and documented. This directly fed into WP2's objective of providing a thorough taxonomy of the various definitions of data in use among different research communities. This amendment was included in the interim management report which was approved by the project's European Project Officer.

Both of these instruments (the interviews and the data mining exercise) were designed to tease a more detailed picture of what model of data underlies computer science research and development. The data mining exercise was used largely at this level of definitions only, while the interviews will sought to delve more deeply into awareness and conceptions of the ethics and responsibilities of aggregating data that is an incomplete record of phenomena, as well as any strategies deployed to address this challenge. Our target was to be in excess of 12 1-hour interviews.

T 2.3 Delivery and initial data preparation/analysis of the data mining and interview results (M 9-11)

The data mining exercise was performed across a corpus of computer science journals and proceedings from “big data” conferences. Interviewees were recruited from the partner networks (in particular the SFI ADAPT Centre, in which the WP leader is an Associate Investigator.)

T 2.4 Analysis of data, write up and editing of reports (M 12-14)

Although data analysis and interview transcription were continuous throughout the phase of Task 2.3, this task will took the final data set and determined the overall conclusions and recommendations that would become a part of the final WP reports.

T 2.5 Integration of final WP results with overall project (M 15)

The final month of the project will be dedicated to the alignment and harmonisation of final results among the WPs, which will be pushed forward by the final project plenary meeting at the start of M 15)

All of the tasks will be led by WP2 leader (TCD). Input of the other partners (KNAW-DANS, FUB and TILDE) will be gathered at the kickoff, midterm and final project meetings as well as through the monthly PMB online conference calls, as well as asynchronously through the circulation of draft results.

2. Literature Review for *WP2.*

2.1. The influence of information studies, archival studies, and documentation studies on how we conceive of and identify data.

The definitions, distinctions, and curatorial liberties granted to the documentationalist as outlined in Suzanne Briet's *What is Documentation* have become integral to archival scholarship and conceptions of what data is, so much so that facets of Briet's documentationalist approach have become enshrined in archival practice, providing prescriptive and unquestioned approaches to documentation that have been automatically translated into the digital sphere and co-opted into digital archival practice. Analogue documentation practice, what Ben Kafka refers to as "the 'charismatic megafauna' of paperwork,"[1] has greatly influenced digital documentation practice, meta-data, and information architecture.

Speaking in relation to projects such as Project Gutenberg and Google Books Rosenberg notes "Some of these resources are set up in ways that generally mimic print formats. They may offer various search features, hyperlinks, reformatting options, accessibility on multiple platforms, and so forth, but, in essence, their purpose is to deliver a readable product similar to that provided by pulp and ink."[2] The longstanding (and long-acknowledged) problems of analogue documentation and analogue forms of metadata have carried over to digital practice, shaping the architecture of digital documentation (and consequently the archives as a whole), and the epistemological implications of these structures, and their effect on how we approach and access stored material (I hesitate to use the term data) is not receiving the attention it deserves.

Briet focuses on the importance of metadata in the form of the catalogue. The centralised role of metadata has been argued as necessary for functionality, for practicality (planning research trips, navigating the archive etc.) and for documenting the contents of the archive, to the point where its role as a mediator between scholar/ user and archive has become standard; metadata is now not only an obligatory intermediary between a scholar and an archive, but a structural facet of the information architecture that mediates the space between user interface and the database:

> Current *catalogues*, retrospective catalogues, and union catalogues are obligatory documentary tools, and they are the practical intermediaries between graphical documents and their users. These catalogues of documents are themselves documents of a secondary degree.[3]

This has transitioned across from analogue to digital and feeds into the information architecture of digital archives. This sets in place a hierarchizing process from the onset,

[1] Ben Kafka, The Demon of Writing: Powers and Failures of Paperwork (Zone Books, 2012), 10.

[2] Daniel Rosenberg, "Data before the Fact,"" Gitelman, *"Raw Data" Is an Oxymoron*, 22.

[3] Suzanne Briet et al., *What Is Documentation?: English Translation of the Classic French Text* (Lanham, Md: Scarecrow Press, 2006), 11.

privileging the facets summarised in the metadata. Again, this goes back to Briet's notion of the documentationalist as curator; and this curatorial control can be seen in the metadata developed for projects such as the SHOAH Visual History Archive.

Digital data management's indebtedness to analogue information science is visible not just in the comparability between print and digital methodologies that mimic print, but in the discipline specific vocabulary used by experts: "Long ago, data managers moved past speaking in narrow technical terminologies, such as 'storage' and 'transmission,' and turned to a more nuanced vocabulary that included 'data preservation,' 'curation,' and 'sharing.' These terms are drawn from the language of library and archival practice; they speak to the arrangement of people and documents that sustain order and meaning within repositories."[4]

Perhaps what's being overlooked here is this very indebtedness to the language of library and archival practice, and to documentationalism; Briet places the discretion and curatorial agency of the documentationalist very high on her list, this grants the documentationalist and archivist or cataloguer an agency and control over the materials is perhaps no longer ideal, and was perhaps never ideal. Granted, it was necessary for the foundation and systemisation of this field of speciality, but perhaps the implications of these practices need to be reconsidered in terms of their incorporation into a digital environment and their use in relation to big data; particularly in relation to data7 of a sensitive nature. There is a similar need to question the curatorial agency of the database structure, its performative influence on the shape of the database, and how this impacts on and shapes the data.

What Briet does do nicely is articulate a prejudice or perceived distinction between humanities and the sciences that seems to be still in place today; namely, that, in her opinion, the humanities are slower or more reluctant than the sciences to adopt innovative methods, and that this reluctance is correspondingly reflected in the documentation habits discernible between the sciences and the humanities:

> Documentation for oneself or for others has appeared in the eyes of many people as '*a cultural technique*' of a new type. *This technique has prospered, first of all, in the area of scientific research*, properly speaking, that is, in the sciences and their applications. The human sciences adopted it more belatedly. [...] In contrast, in the fields of the human sciences, documentation proceeds by accumulation: literature, history, philosophy, law, economics, and the history of the sciences itself are tributaries of the past. Erudition is conservative. Science is revolutionary. *The evolution of human knowledge* is a permanent compromise between two mental attitudes. Invention and explanation, reflection and hypothesis divide the field of thought.[5]

The drawbacks and shortcomings of this approach have long been acknowledged, even by Briet herself when she quoted Samuel C. Bradford:

[4] David Ribes and Steven J. Jackson, "Data Bite Man: The Work of Sustaining a Long-Term Study" Gitelman, *"Raw Data" Is an Oxymoron*, 147.

[5] Ibid., 13, emphasis in original.

> Moreover, a detailed study of the work of analytical journals led him [Samuel C. Bradford] to the conclusion that, in principle, two-thirds of the collections of specialized documentation agencies did not directly relate to the profile of the agency; and that nonetheless, the tonality of the documentation of interest to the specialty couldn't be found anywhere.[6]

Again, Briet's analysis below of the two distinct tendencies discernible in documentary techniques is still applicable today:

> [D]ocumentary techniques very clearly mark two distinct tendencies. The first is towards an always increasingly abstract and algebraic schematization of documentary elements (catalogs, codes, perforations, classifications through conventionally agreed upon marks). The second is toward a massive extension of 'substitutes for lived experience' (photos, films, television, audio records, radio broadcasting). [...] What words fail to communicate, image and sound try to deliver to all.[7]

Briet's *What is Documentation?* clearly shows that analogue documentation and archival practice is hierarchical in nature. Briet grants interpretative/ curatorial authority to the documentalists; and this curatorial authority has transferred/ is analogous to the authority of the researcher to select/ hierarchise material within the confines of a given research project. But when it comes to digitising materials, is it enough to replicate this already problematic process? Should digitising projects not try to undo the hierarchies/restrictions imposed within the analogue archive? How can this be done apropos the recording of material?

2.1.1 The role and influence of metadata in a digital environment

Classification systems come in the form of taxonomies, metadata, ontologies, controlled vocabularies, folksonomies, and crosswalks. Research Classification Systems provide high level classification with three to four facets max. These are often developed and implemented at a National Level and are thus country specific with the result that they are referred to as National Research Classification Systems. As a consequence, the metadata is not granular. Classification is for the purpose of research evaluation and the most important and influential of these is the Australia and New Zealand classification systems.

In terms of a history this can be traced to 1963 with the OECD's Frascati Manual (*The Proposed Standard Practice for Surveys of Research and Experimental Development*) and later the Fields of Science (FOS) classifications. This was revised and became the *Revised Field of Science and Technology (FOS) classification in the Frascati Manual*, frequently referred to as the Frascati Fields of Science. They provide the user or institution with a choice of classifications (which are maintained by UNESCO). The Frascati Fields of Science (the revised Frascati Manual) was taken and adapted by the Australia and New Zealand governments and adapted to develop the ANZSRC (Australian and New Zealand Standard Research Classification):

[6] Ibid., 14.

[7] Ibid., 31.

> ANZSRC is the collective name for a set of three related classifications developed for use in the measurement and analysis of research and experimental development (R&D) undertaken in Australia and New Zealand. The three constituent classifications included in the ANZSRC are: Type of Activity (TOA), Fields of Research (FOR), and Socio-economic Objective (SEO).
>
> ANZSRC replaces the Australian Standard Research Classification (ASRC 1998) and introduces a new framework for measuring research and development activity in New Zealand.
>
> The use of the three constituent classifications in the ANZSRC ensures that R&D statistics collected are useful to governments, educational institutions, international organisations, scientific, professional or business organisations, business enterprises, community groups and private individuals in Australia and New Zealand.[8]

An analogous framework in operation in the UK is the REF (Research Excellence Framework), which was developed for the purpose of research evaluation.[9] Similar classification systems are under development in Belgium, Ireland, and elsewhere.

In terms of European Research we have the European Current Research Information Systems (euroCRIS) and the Common European Research Information Format (CERIF):[10]

> The mission of euroCRIS is to promote cooperation within and share knowledge among the research information community and interoperability of research information through CERIF, the Common European Research Information Format. Areas of interest also cover research databases, CRIS related data like scientific datasets, (open access) institutional repositories, as well as data access and exchange mechanisms, standards and guidelines and best practice for CRIS.[11]

Many of these research classification systems are for research evaluation and quality judgements, they classify the “type of research” and aim to assess or provide a measure of its quality. This is the case with the ANZSRC and with the UK’s NRC (REF). The fact that NRC’s involve high level—and not granular— classification is seen as both an advantage and a disadvantage, because it flattens out the data and makes the granular more difficult to identify and access. Aside from research classification, organisations such as ISCED (International Standard Classification of Education), which is maintained by UNESCO, provides a framework for classifying education on an international level. Additional examples of research classification systems that are incorporated into library databases are SCOPUS, which is an abstract and citation database, so again is for research classification and evaluation purposes and to measure research impact. In the case of libraries, the International Federation of Library Associations (IFLA) are the international body responsible

[8] http://www.abs.gov.au/Ausstats/abs@.nsf/Latestproducts/1297.0Main%20Features32008
[9] http://www.ref.ac.uk/about/
[10] http://www.eurocris.org/
[11] http://www.eurocris.org/what-eurocris

for representing library and information services. They are based in The Hague and their headquarters are to be found in the Royal Library, the National Library of the Netherlands.

The Dewey Decimal Classification (DDC) system continued to be widely used internationally. The Library of Congress in D.C. have taken over management of the Dewey system, and The Dewey Programme at the Library of Congress works to "to develop, apply, and assist in the use of the Dewey Decimal Classification (DDC)."[12] Despite this the Library of Congress have their own classification system, creatively titled the Library of Congress Classification System (LC). Both the DDC and LC make use of controlled vocabularies. In addition to the controlled vocabulary thesauri (to be discussed presently) we have projects such as "Metadata Image Library Exploitation" (MILE) which focus on categorising and classifying images, embedding rights information and developing interoperable metadata for the digital images of fine artworks or other artistic works. Along with MILE there is the ARROW system which stands for "Accessible Registries of Rights Information and Orphan Works towards Europeana." ARROW works to "create registries of orphan and out-of-print works together with a network of rights management mechanisms."[13]

What is of interest to this study in terms of data complexity and classification systems are the controlled vocabularies adopted by these classification systems. These classification systems extend to the classification of the visual with VRA Core (which is based on Dublin Core) being a data standard for the description of visual cultural artefacts and their documenting images; VRA Core provides schemas, description and tagging protocols, and category guides. We also have the Getty Image Classification system which has subsections devoted specifically to the arts and humanities such as the AAT (Arts and Architecture Thesaurus), the ULAN (Union List of Artist Names), and CONA (Cultural Object Name Authority). These promote very definite views on classification, providing structured terminologies with the aim of making objects discoverable through standardization of classification. Again, this allows for high level accessibility, but not granularity or idiosyncrasy. Furthermore, these set vocabularies provide a standardised approach to the indeterminate of unknown using words such as "circa" for uncertain dates and terms such as "anonymous" for uncertainty regarding authorship.

Alongside controlled vocabularies governed by set thesauri, there are also locally used classifications and folksonomies, which feature locally defined classification systems or, in the case of folksonomies, user contributed keywords or user-generated tagging mechanisms. Together these provide a matrix of metadata, controlled vocabulary, taxonomies and classification systems that make up the cataloguing and metadata rules that have been adopted by GLAM institutions.

[12] https://www.loc.gov/aba/dewey/ (Accessed 4/7/17)

[13] Estelle Derclaye, *Copyright and Cultural Heritage: Preservation and Access to Works in a Digital World* (Edward Elgar Publishing, 2010), 205.

2.1.2. The transition from analogue to digital environments.

Moving on from documentationalist approaches to analogue material to the computerised database itself. The recognition and justification of the need for databases in the first place:

> a database and, hence, the very genre of computational representation exists, first of all, to manage scale. Secondly [...] a relational database, by definition, functions by virtue of the relations or cross-connections between the data in the database. As such, the database can give rise to infinitely many search queries and thereby allow innumerable combinations that identify larger thematic issues, reveal patterns and structures, and create new associations between experiences that may not otherwise be considered together. [14]

Presner is here being somewhat hyperbolic, especially given the limitations imposed on the SHOAH archive by a keyword search functionality that does not also accommodate user contributions and therefore the capacity to create these "new associations."[15] In the analogue age, scholars obtained information by means of personally directed research, with an increasing reliance on catalogues, abstracts, and reports:

> As ever, the scholar obtains information by his personal relations, by his readings, and by the bibliography he finds there [in the library]. But more and more, he becomes informed by abstracts and by reports.[16]

Briet's question then "Is the scholar confident of having the power to locate *the entirety of that documentation* which interests him?"[17] can be answered assuredly with a No, when we are speaking of digital big data archives whose accessibility is mediated by metadata structures that delimit the search potential/ accessibility of the archive. Further still, the archive may be there in its entirety, but it is only accessible through specific search programmes, only certain information is confined to the realm of metadata, and there is a lack of user awareness of just how complete the resources in question are.

Perhaps then it is our very understanding of data itself that is contributing to these problems. Borgman cites the need to

> address the processes by which something becomes data. How do individuals, teams, and communities create, select, or use data? What factors in these decisions are associated with the data per se? Which are associated with the research questions or methods? Which are functions of how data are represented? How do these considerations vary by field, discipline, and research problem? How do they vary by relationship to the data, from creator to curator?[18]

---

[14] Presner, in Fogu, Claudio, Kansteiner, Wulf, and Presner, Todd, *Probing the Ethics of Holocaust Culture*.

[15] Presner, in ibid.

[16] Briet et al., What Is Documentation?, 16.

[17] Ibid., emphasis in original.

[18] Borgman, "Big Data, Little Data, No Data," 17–18.

We know, for example, that "Data are not pure or natural objects with an essence of their own." [19] We also know that "They exist in a context, taking on meaning from that context and from the perspective of the beholder. The degree to which those contexts and meanings can be represented influences the transferability of the data."[20] But how much of this context is being redacted/ modified / hidden by the information architecture? In the transition from analogue to digital environments metadata takes on an additional level of import, being the determining factor in whether or not an item viewed within the indexes (etc.) of a digital archive are of relevance to the researcher. To what extent it the so-called "raw data" shaped by the database and contexts imported onto it by the metadata in and of itself? And what of the contexts that accompanied the data within its "original" "proto-data" or "native" environment, how are these transferred to and catered for in the database? Again this relates to the "*Surrogates versus Full Content*"[21] debate. As surrogate material such as catalogues and abstracts increasingly replaces or stands in for full content, data or potential data becomes hidden; or, if we want to eliminate the implication that this hiddenness is intentional, the data becomes latent in the archive. Present but not there; existing but not manifest.

Many (if not all) of the definitions that will be outlined throughout section 2.2 involve "native data" that has undergone extensive *cleaning* or *scrubbing*. The phrase "scrubbing data" is taken from economics where it signifies a

> set of elaborate procedures [...] which Wall Street analysts now call "data scrubbing" [...] process of commensuration which could also be called "cleansing" or "amending," involved removing incorrect or inconvenient elements from the available data, supplying missing information, and formatting it so that it fit with other data.[22]

This cleaning of scrubbing is often undocumented, and this is particularly important to the KPLEX project. The technique of cleaning data so as to produce more refined data is not systematised, rather, as Borgman notes, it is presented as an almost esoteric process: "Cleaning these kinds of data is as much art as science, requiring considerable methodological and statistical expertise (Babbie 2013; Shadish, Cook, and Campbell 2002)."[23] Cleaning data implies that that which is removed is "dirty" or unwanted; this might be the case for the particular study at hand, but if has the potential to prejudice users into thinking the cleaner data is somehow better, that the material that was removed is of less value *in general* (as opposed it simply to being external to the remit of the study the data was cleaned for, for example) and to subsequently disregard the removed "native" materials. If data is of speculative value these once discarded facets could become of import at any time in the future, but this is only viable if they are available. As Raley notes, "Data cannot 'spoil' because it is now speculatively, rather than statistically, calculated."[24] It can, however, become hidden, or be rendered latent in the archive as a result of the information

[19] Ibid., 18.

[20] Ibid.

[21] Ibid., 167.

[22] Brine and Poovey, "From Measuring Data to Quantifying Expectations: A Late Nineteenth-Century Effort to Marry Economic Theory and Data," in Gitelman, *"Raw Data" Is an Oxymoron*, 70.

[23] Borgman, "Big Data, Little Data, No Data," 27.

[24] Rita Raley, "Dataveillance and Countervailance" Gitelman, *"Raw Data" Is an Oxymoron*, 124.

architecture employed to facilitate its transition to a digital environment and its inclusion within a digital archive. Raley observes (though here perhaps the term native data would be more appropriate): “Raw data is the material for informational patterns still to come, its value unknown or uncertain until it is converted into the currency of information.”[25] So down the line, if data is scrubbed and left disengaged from its native context, this has the potential to influence the DIKW progression, delimiting that which can be sourced as data and transferred on to information, knowledge and wisdom.

Brine and Poovey capture this paradox in their description of the scrubbing process as one “of elaboration and obfuscation”[26]; it elaborates certain facets of the material, and obfuscates others. Borgman, as always, makes explicit that each and every decision made in the handling and rendering of data has consequences:

> Decisions about how to handle missing data, impute missing values, remove outliers, transform variables, and perform other common data cleaning and analysis steps may be minimally documented. These decisions have a profound impact on findings, interpretation, reuse, and replication (Blocker and Meng 2013; Meng 2011)[27]

In the sciences cleaning/ scrubbing of data is considered standard practice; so much so that it is often taken as a given and, as noted above, minimally documented.[28] While the knowledge that data is always already “cleaned” or “scrubbed” is implicit in disciplines such as economics or sociology, not so in other disciplines: “even an economist who wanted to do empirical work [...] always carried over the assumptions implicit in the ways these data were collected and presented in the first place, even as his use of them effaced their historical and conventional nature.”[29]

Examples on the difficulties of scrubbing data (in this case in the form of an early modern maths problem):

> The ostensible goal is to find a way to cut away everything that is not mathematics and thereby leave the mathematics scoured and ready for further investigation. The methodological irresponsibility of such an effort appears immediately, since the decision about what is mathematics and what is not can only proceed via a modern sense of the distinction. Some examples of current arithmetical pedagogy show that we, perhaps more than people in other eras, are particularly keen to segregate mathematics from anything that doesn’t fit our sense of what mathematics should be.[30]

---

[25] Rita Raley, “Dataveillance and Countervailance” ibid., 123.

[26] Brine and Poovey, “From Measuring Data to Quantifying Expectations: A Late Nineteenth-Century Effort to Marry Economic Theory and Data,” in ibid., 73.

[27] Borgman, “Big Data, Little Data, No Data,” 27.

[28] See ch 8 Raw Data is Oxymoron re the cleaning that takes place in long term projects.

[29] Brine and Poovey, “From Measuring Data to Quantifying Expectations: A Late Nineteenth-Century Effort to Marry Economic Theory and Data,” in Gitelman, *“Raw Data” Is an Oxymoron*, 71–72.

[30] Williams, “Procrustean Marxism and Subjective Rigor: Early Modern Arithmetic and Its Readers,” in," in ibid., 47.

William's argues here that native context is essential, and an abstraction from context is described as irresponsible, an act that forces "a teleology that is not there": "To ignore these qualities [of early modern arithmetic] and focus instead on those aspects that are similar to our own mathematics is to force a teleology that is not there."[31] A further example, this time from economics:

> Not only is normative economic data never raw, then, in the sense of being uninterpreted, but also the form that makes data suitable for economists' use carries with it assumptions about the quantification and value that now go unnoticed and unremarked.[32]

In the humanities, cleaning of data does not receive as much (or any) attention or acknowledgement. Contradictions and confusions are rampant across humanities disciplines with respect to what data is, how data becomes data (whether it needs to be cleaned or scrubbed, whether the fact that this scrubbing takes place is or is not implicit to the discipline) and whether original context, abstraction, or recontextualising are integral functions for the treatment of data.

Gruber Garvey's essay on *American Slavery As It Is*, for example, highlights *American Slavery As It Is* as a project that "helped to create the modern concept of information, by isolating and recontextualizing data found in print."[33] Here data of a very specific kind is used to create information, but earlier on the same page, information is somehow "pried loose" from "abstracted" advertisements that have been "accumulated, aggregated en masse"[34]: "It was this work of trimming, sifting, and aggregating the material that recreated it as a database and not just a collection of anecdotes: This work allowed for its recontextualization and analysis."[35] This process seems rather to involve the creation and curation of proxy data. Nevertheless, it makes clear that the scrubbing and obscuring of material is nothing new; but that this activity can even go so far as to "imagine" data, in other words the process of displacement is integral to the proto-data becoming data proper: "Not only have their collective project imagined data by dint of its displacement from its original Southern contexts, their project had also depended on an additional displacement of those same sources from the commercial and the domestic sphere."[36]

At the same time, the counter-argument is also compelling; Markus Krajewski in "Paper as Passion: Niklas Luhmann and His Card Index": "What good it the most meticulous transcript if it cannot be brought into productive relationship with other entries? What good are page-long excerpts if they do not inscribe themselves into a network of pre-forming cross-references?"[37] A similar point also voiced by Rita Raley: "Even if one were to accept the fiction of the universal database managed by a single authority, the fundamental problem of

[31] Williams, "Procrustean Marxism and Subjective Rigor: Early Modern Arithmetic and Its Readers," in ibid., 52.

[32] Brine and Poovey, "From Measuring Data to Quantifying Expectations: A Late Nineteenth-Century Effort to Marry Economic Theory and Data," in ibid., 61.

[33] Ellen Gruber Garvey, "'facts and FACTS:' Abolitionists' Database Innovations," ibid., 91.

[34] Ellen Gruber Garvey, "'facts and FACTS:' Abolitionists' Database Innovations," ibid.

[35] Ellen Gruber Garvey, "'facts and FACTS:' Abolitionists' Database Innovations," ibid., 96–97.

[36] Ellen Gruber Garvey, "'facts and FACTS:' Abolitionists' Database Innovations," ibid., 98.

[37] Markus Krajewski, "Paper as Passion: Niklas Luhmann and His Card Index" ibid., 114.

meaningfully, and predictably, parsing that archive remains."[38] Again this returns us to the "*Surrogates versus Full Content*"[39] debate. Briet in contrast outlines the necessity for the scholar to have full accessibility: "Is the scholar confident of having the power to locate *the entirety of that documentation* which interests him?"[40] The necessity for totality is particularly rampant in the sciences (but also in the humanities too, surely; specifically for a discipline like literature wherein one could not imagine a scholar confidently writing on a text without having read it in its entirety). A variant on this is the necessity in the sciences in particular for data to be comparable across time: "Data must be comparable across time and sufficiently well described so as to facilitate integration with other data."[41]

Context is sometimes considered relevant, sometimes discarded, and sometimes replaced with another context deemed more appropriate to bring out the latent information in the "data." Despite the enormity of the *American Slavery As It Is Today* project, the scale of proliferation in contemporary big data environs has now immeasurably increased. Question: This modified data, can it be returned to its "native" state? Should data's level/ degree of treatment be noted in a manner akin to the way it is in the NASA EOS DIS table? How do we maintain this capacity to both extract material from its original source and situate material in its native context, to see it both in context and in its recontextualised context?

Distinctions and different practices exist regarding how data is treated in digital environments—or how researchers think data *should be* treated—in different factions of the sciences; for example in mathematics (Williams's chapter) and here in the sciences in ecological research. Williams emphasizes the need to maintain awareness of the distinctions between then/ now and them/us,[42] whereas Ribes and Jackson identify temporality as a challenge to the maintenance and sustenance of a long-term study.[43] There are also different approaches within these same disciplines, as for example when Williams's acknowledges that within the field of mathematics different scholars will identify different facets of the examples he highlights as important or useful. The question remains however over who gets to make the decision over whether the original context is relevant or not. And, as seemed to be the case in Krajewski's account of Card Indexing wherein the indices supplanted the original texts, so too in a digital big data environment the metadata, the description of the document, has the capacity to supplant and replace the document.

A central paradox of the approach to the transition from analogue to digital environments in the humanities is captured in the following excerpt:

> The humanities are experiencing as much of a data deluge as other fields but are less likely to call it by that name. To cope with the scale of content available to them, scholars are borrowing technologies and methods from related fields and developing

[38] Rita Raley, "Dataveillance and Countervailance" ibid., 129.

[39] Borgman, "Big Data, Little Data, No Data," 167.

[40] Briet et al., *What Is Documentation?*, 16, emphasis in original.

[41] David Ribes and Steven J. Jackson, "Data Bite Man: The Work of Sustaining a Long-Term Study" Gitelman, *"Raw Data" Is an Oxymoron*, 147.

[42] Williams, "Procrusteam Marxism and Subjective Rigor: Early Modern Arithmetic and Its Readers," in ibid., 42.

[43] David Ribes and Steven J. Jackson, "Data Bite Man: The Work of Sustaining a Long-Term Study" ibid., 147.

> some of their own. In the process, they are beginning to think in terms of data, metadata, standards, interoperability, and sustainability. Borrowing techniques usually requires borrowing concepts. They too need ways to represent information in computationally contractable forms[44]

We have seen the level of preparedness, or at least the willingness to classify and categorise data, that exists within the sciences; this is not in evidence to the same degree (or even at all) in the humanities. As Borgman observes in the above passage, scholars are even reluctant to use the term data, and among those that do, there is a discernible tendency towards adopting concepts without adapting them to suit the unique teleological research environments of contemporary humanities research. Borgman acknowledges that in "*data scholarship*" "the set of relationships in the humanities is particularly complex and not always a comfortable framework for the scholars involved."[45]

Masson acknowledges Drucker's contribution to our understanding of the indebtedness of these infrastructures "to the disciplines from which they derive" and develops on this by noting that

> The same, one might add, applies to tools for data scraping, and for the cleaning, sorting or otherwise processing of collected data. For digital humanists, this is particularly relevant, as the tools they use are rarely purpose-produced (or if they are, then they tend to refashion tools that were designed to serve the needs of other disciplines).[46]

According to Ribes and Jackson, from a scientific perspective these invisible infrastructures are "Technicians, robots, and cooling systems are increasingly hidden in the clouds of computing, laboring to preserve the data of the earth sciences and, agnostically, those of many others."[47] Together this makes up a

> long chain of coordinated action that stretches back directly to a multitude of local sites and operations through which data in their 'raw' form get mined, minted, and produced. What remain at repositories are the distilled products of these field sites; behind these centers lie an even more occluded set of activities that produce those data themselves.[48]

But there are other invisible infrastructures that Robes and Jackson overlook. These come in the form of the algorithms that make data, and particularly big data, available and navigable in a digital environment. William Uricchio:

---

[44] Borgman, "Big Data, Little Data, No Data," 161–62.

[45] Ibid., 162.

[46] Eef Masson, "Humanistic Data Research An Encounter between Epistemic Traditions," in Mirko Tobias Schäfer and Karin van Es, The Datafied Society. Studying Culture through Data, 29.

[47] David Ribes and Steven J. Jackson, "Data Bite Man: The Work of Sustaining a Long-Term Study" Gitelman, *"Raw Data" Is an Oxymoron*, 152.

[48] David Ribes and Steven J. Jackson, "Data Bite Man: The Work of Sustaining a Long-Term Study" ibid.

> the algorithm is a talisman, radiating an aura of computer-confirmed objectivity, even though the programming parameters and data construction reveal deeply human prejudices. The bottom line here is that decisions regarding what will and will not be produced are now often based on data of unknown quality (What do they actually represent? How were they gathered?) that are fed into algorithms modelled in unknown ways (What constitutes success?).[49]

The crossover between the sciences and humanities appears to be primarily one directional, with methodologies, techniques and software from computer science (in particular) infiltrating/ being incorporated into humanities research, with the exception of course of long-lasting endeavours such as WordNET and GermaNet, or the development of ontologies by humanists. These methodologies are being incorporated (sometimes with minimal adaptation) into these new research environments, with the epistemological implications of same left under-examined. Similarly, the core centrality of uncertainty and ambiguity to humanities research is at odds with the unambiguity necessary for computer technologies to function. There is a need to develop approaches to data and big data that are humanities-specific and sensitive to the uniqueness and idiosyncracies of humanities research:

> Rather than import questions and methods from the hard sciences, we must develop our own approaches and sensitivities in working with data that will reflect the humanities' traditions.[50]

There are growing concerns also over the ethics of facets of this research, thus far mainly concerned with relationship between form and content (in the case of Presner), but also some expressions of anxiety over the lack of scholarly investigation into the epistemological implication of the software techniques being co-opted into DH practice. This is where I think Presner is on point regarding the ethical potential of the database: it allows for scale and appreciation (if that is the right word) of scope; this stands in contract with analogue historical practice which focuses on single case-studies or singular accounts; instead here we can attempt to take in the fullness of the archive, and appreciate that only insofar as it is a partial simulacra of records pertaining to the Holocaust: "computational or algorithmic analysis can be ethical precisely because it takes into account the totality of the archive insofar as all the indexed data related to the narrative of every survivor is part of the analysis."[51]

There is, however, a growing awareness that this information needs to be shared and incorporated into our discussions of DH structures:

> Data capture (starting with the transfer of the video tape to digital formats and cataloguing) to the storage of data (both the videos themselves and the indexing

[49] William Uricchio, "Data, Culture and the Ambivalence of Algorithms," Mirko Tobias Schäfer and Karin van Es, The Datafied Society. Studying Culture through Data, 131–132.

[50] Ibid., 15.

[51] Presner, in Fogu, Claudio, Kansteiner, Wulf, and Presner, Todd, *Probing the Ethics of Holocaust Culture*.

> server that knows where all the catalogue metadata is) and, finally, the interface to play, search for, and distribute data and its related content.[52]
>
> The
> realm of information architecture between the user interface and the server storage– in other words, the metadata, the data structures, and the database. It is precisely here that we see a fundamental disassociation of the presentation of the content (that is, the testimonies and the interface to watch them) from the information architecture, database, and metadata scaffolding that lies behind the content.[53]

Presner's concern is with the disassociation of content and form. And he wonders whether content and form could be integrated, citing it as a problematic factor for the SHOAH VHA, but also in other digital archives: "Such a dissociation is not unique to the VHA but bespeaks a common practice in digital library systems and computation more generally, stretching back to Claude Shannon's theory of information as content neutral."[54]

Presner argues "There is no reason, then, why the realm of information architecture, data structures, and databases should be considered apart from the realm of ethics and the subjective, contingent, meaning making, interpretative practices at the heart of the humanities." And argues against "completely separating content from information architecture, of priviledging dis-ambiguated data ontologies over probabilistic knowledge, potentialities of figuration, and interpretative heterogeneity. But computational representation does not have to be this way if it is guided by an ethics of the algorithm."[55]

What Presner's overlooking or taking for granted is this idea of "information as content neutral." We would all now likely accept that *'Raw Data' is an Oxymoron*, what is not yet as widely acknowledged is that neutrality in computation and coding is also an oxymoron. Everything is the product of some degree of human interference or has in some way been curated or constructed. The VHA "knowledge model, [...] is fundamentally aimed at the transformation and dis-ambiguation of narrative into data that makes it amenable to computational processing and structured logic. It is a process that can be called 'de-figuration'—precisely because it evacuates all traces of the figurative in its literalism."[56] The problem Presner identifies is the disambiguation, the flattening of narrative and its defiguration into data. He posits this "double-blind" as an ethical dilemma arguing that somehow the content and form should intermingle or be sensitive to each other. But it is not the only ethical problem here, there are others. Bowker also cites this "flattening" effect: "We have flattened both the social and the natural into a single world so that there are no human actors and natural entities but only agents (speaking computationally) or actants (speaking semiotically) that share precisely the same features."[57]

---

[52] Presner, in ibid.

[53] Presner, in ibid.

[54] Presner, in ibid.

[55] Presner, in ibid.

[56] Presner, in ibid.

[57] Geoffrey C. Bowker, "Data Flakes: An Afterword to 'Raw Data' is an Oxymoron," " Gitelman, *"Raw Data" Is an Oxymoron*, 170.

But do we not also encounter this same content/ form dilemma in narrative? Couldn't we also argue, for example, that narrativised accounts of the Holocaust defigure and disambiguate the reality of the event itself? Or that, in the case of Ribes and Jackson's representative ecological study wherein "the natural world is translated, step by step, from flowing streams to ordered rows of well-described digital data, readily available for use in science,"[58] the same process takes place? Isn't this what data is in the first place? Not the event itself in its totality, but the documentation, the traces (etc.) of the event? As observed later in section 2.2, Borgman describes data as "data are representations of observations, objects, or other entities used as evidence of phenomena for the purpose of research or scholarship."[59]

The following excerpt makes explicit the extent to which Presner overlooks the epistemological implications of computational software:

> A mode of organizing information characterised by the 'separation of content from material instantiation ... [such that] the content management at the source and consumption management at the terminus [are] double-blind to each other.' In other words, the content of the testimonies knows nothing of the information architecture, and the information architecture knows nothing of the testimonies. In this sense, the database is simply an empty, neutral bucket to put content in, and the goal of the information system is to transmit this content as noiselessly as possible to a receiver or listener.[60]

That said, it is perhaps unfair to cite this as an oversight given that "the tools they [Digital Humanists] use are rarely purpose-produced (or if they are, then they tend to refashion tools that were designed to serve the needs of other disciplines."[61] In addition then to sharing material such as the architecture of digital libraries etc, the practices adopted must be interrogated.

There is a sense that humanities scholars are unprepared/ ill equipped/ unfamiliar with the technologies and methodologies used in the DH, technologies that have been adapted from science and computer technology. They are thus more likely to blame/ ignore/ discard/ dismiss the epistemological fallout brought on by the digitizing process (etc.) than adapt the processes to suit the materials being represented or digitized. Alternatively, and this is the point noted by Borgman, they may blindly adopt these processes, without considering the epistemological implications of doing so, or without considering that these techniques, when inducted into a humanities environment, must necessarily be examined from an epistemological perspective and become more fluent parlance and more integrated into epistemological studies of humanities research. This is further complicated by the fact that within computer science the act of "borrowing" or "repurposing" software is itself common practice:

---

[58] David Ribes and Steven J. Jackson, "Data Bite Man: The Work of Sustaining a Long-Term Study" ibid., 147.

[59] Ibid., 28.

[60] Presner, in Fogu, Claudio, Kansteiner, Wulf, and Presner, Todd, Probing the Ethics of Holocaust Culture.

[61] Eef Masson, "Humanistic Data Research An Encounter between Epistemic Traditions," in Mirko Tobias Schäfer and Karin van Es, The Datafied Society. Studying Culture through Data, 29.

> Building working software is not just about writing lines of code. It also involves orchestrating different pieces of existing software (such as databases, networking libraries, and visualization frameworks) to work together in what is called a 'technology stack'. Programmers reuse existing software as much as possible to minimize the effort needed to produce new software, knowing that prior programmers will have already encountered problems and devised and tested solutions that would otherwise have to be done anew. This recombinatory aspect is also part of what could be called the programmer's poetics or particular style of working.[62]

2.1.3 The role of narrative; the role of code and programming.

A central facet of this evolving discomfiture comes in the form of our difficulties outlining the relationship between narrative and data. In the interplay between analogue and digital, different factions emerge regarding the relationship between data and narrative, with narrative and data variously presented as being anathematic, antagonistic, or symbiotic, with data presented as something one can be either "for" or "against" and with distinct preferences for one or the other (either narrative or data) being shown on a discipline specific and researcher specific level: "what would it mean to be against data? What would it mean, for that matter, to be for data?"[63]

Narrative and data are presented as antithetical by Manivoch (2002) and Rosenthal (2017). Lev Manovich calls narrative and database "natural enemies."[64] Rosenthal observes how "the novel form itself is consistently adept at expressing the discomfort that data can produce: the uncertainty in the face of a central part of modern existence that seems to resist being brought forward into understanding."[65] Rosenthal's engagement is more open to and acknowledging of the established relationship between narrative and data in literature, and the emergent relationship between narrative and data in digital environments, and the impact this has on both traditional literary criticism, and in the form of emergent data-driven criticism: "Narrative and data play off against each other, informing each other and broadening our understanding of each."[66]

A slight variant on the above is the opinion voice by Presner: that narrative begats data, and data begat narrative, though Rosenthal also presents data and narrative as operating in this systematised manner.[67] In the context of an analogue to digital environment this means there is a seemingly irreconcilable transference of medium in either direction; one that has ethical and epistemological implications that go beyond the remit of Presner's study:

---

[62] Joris J. van Zundert, Smiljana Antonijević, Tara L. Andrews, "Black Boxes and True Color—A Rhetoric of Scholarly Code," draft chapter, (unsure how to footnote this), 14.

[63] Rosenthal, "Introduction," 1.

[64] Lev Manovich, The Language of New Media, 2002, 228.

[65] Rosenthal, "Introduction," 2.

[66] Ibid., 4.

[67] "Yet the narrative relies for its coherence on our unexamined belief that a preexisting series of events underlies it. While data depends on a sense of irreducibility, narrative relies on a fiction that it is a retelling of something more objective. [...] The coherence of the novel form, then, depends on making us believe that there is something more fundamental than narrative." Ibid., 2–3.

> The effect is to turn the narrative into data amenable to computational processing. Significantly, this process is exactly the opposite of what historians usually do, namely to create narratives from data by employing source material, evidence, and established facts into a narrative.[68]

Kathryn Hayles presents an alternative argument, arguing that data and narrative are symbiotic and should be seen as "natural symbionts."[69] And while the relationship between narrative and data has received much attention from philosophers of science, being variously presented as antagonistic, antithetical or even symbiotic; existing in relationships that can pose ethical and epistemological challenges for the researcher or software engineer (Presner 2017, Rosenthal 2015, Hayles 2007, Manovich 2002), less has been said about how this relationship is perceived among the computer science community.

Narrative plausibility depends on what Rosenthal refers to as a "belief that a preexisting series of events underlies it."[70] The load-bearing *data* that underlies these narratives are considered "more fundamental than narrative."[71] So narrative needs data to render it plausible, be that in a fictional or nonfictional environment. In turn, data needs narrative, because without narrative the average reader of data cannot assess or recognize the plausibility of the data—what the data *means*, what it is *saying*. Data—when understood as a sort of untouched, "objective" input—is considered independent of narrative, with narrative in turn generally understood as the subjective interpretation of data; the stories we tell about data.

Data and narrative are presented by some as being irreconcilable or antithetical (in the context of information architecture). Manovich presents them as "natural enemies."[72] Presner situates database and narrative as being at odds or somehow irreconcilable:

> this is because databases are not narratives [...] they are formed from data (such as keywords) arranges in relational tables which can be queried, sorted, and viewed in relation to tables of other data. The relationships are foremost paradigmatic or associative relations [...] since they involve rules that govern the selection or substitutability of terms, rather than the syntagmatic, or combinatory elements, that give rise to narrative. Database queries are, by definition, algorithms to select data according to a set of parameters.[73]

Are narrative and the database really at odds in this way? Is there a way to reconcile narrative and the database? How can one facilitate and elucidate the other best in a digital environment?

---

[68] Presner, in Fogu, Claudio, Kansteiner, Wulf, and Presner, Todd, Probing the Ethics of Holocaust Culture.

[69] N. Katherine Hayles, "Narrative and Database: Natural Symbionts," PMLA 122, no. 5 (2007): 1603.

[70] Jesse Rosenthal, "Introduction: 'Narrative against Data,'" *Genre* 50, no. 1 (April 1, 2017), doi:10.1215/00166928-3761312.

[71] Jesse Rosenthal, "Introduction: 'Narrative against Data,'" *Genre* 50, no. 1 (April 1, 2017), doi:10.1215/00166928-3761312.

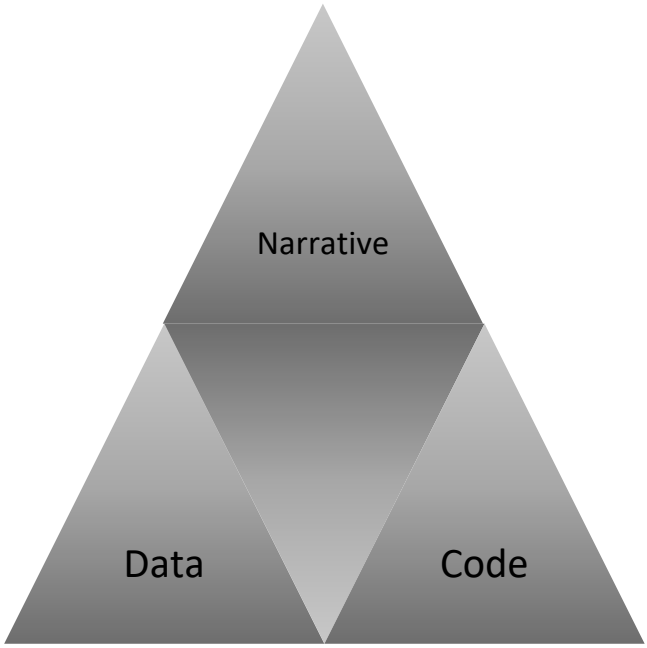
[72] Manovich, *The Language of New Media*, 2002, 228.

[73] Presner, in Fogu, Claudio, Kansteiner, Wulf, and Presner, Todd, *Probing the Ethics of Holocaust Culture*.

In addition, David Ribes and Stephen Jackson recount the process wherein "the natural world is translated, step by step, from flowing streams to ordered rows of well-described digital data, readily available for use in science."[74] This in itself (the creation and maintenance of long-term data stream) can be read as a narrative, just a narrative of a different genre. Perhaps it is time to stop thinking of data and narrative as being at odds with each other; perhaps (as will be suggested presently in section 2.1.4), code is the facet that can enter this dyad and facilitate better integration?

2.1.4 The role of code and programming.

This narrative/ data dyad has received significant attention, but a third facet perhaps needs to be incorporated into our discussions here, one that acknowledged the role of code as a language, and thus a carrier and conveyer of narrative, a shaper, exploiter, user and instrumentaliser, and manipulator of data. Such an approach necessarily involves acknowledging the role of the algorithm and the relationships that exist between narrative and code, and between data and code. There are some fundamental misconceptions about code that need to be redressed if the KPLEX project is to succeed in reconceptualising data so that rich data can be facilitated in a big data environment. This would create a tripartite triadic relationship along the lines of that represented in the following diagram:



Code is not objective but rather is culturally shaped and can even be said to have a "poetics" or "style" that is user specific.[75] Code facilitates both narrative and data within digital environments yet code is largely unacknowledged and invisible, it facilitates these

[74] David Ribes and Steven J. Jackson, "Data Bite Man: The Work of Sustaining a Long-Term Study" Gitelman, *"Raw Data" Is an Oxymoron*, 147.

[75] Joris J. van Zundert, Smiljana Antonijević, Tara L. Andrews, "Black Boxes and True Color—A Rhetoric of Scholarly Code," draft chapter, (unsure how to footnote this), 14.

environments and yet, as van Zundert, Antonijević and Andrews note, "codework remains epistemologically and methodologically unexamined."[76] Also Rieder and Rôhle: "The singular focus on code may detract from what is actually coded."[77] There are rampant misconceptions surrounding the importance of code, about its constituents and the implications of using it. Code is not objective, but rather "codework is necessarily shaped by its social context, a feat made all the more complex considering that "borrowing" or "repurposing" is a common practice in programming. The context of code may positively or negatively influence the attitude and perception that both coders and other scholars hold towards their work."[78] It is possible to reverse this one-directional relationship, by using text to illuminate code, as opposed to subordinating code to text: "Just as code enhances text making it amenable to methodological and epistemological approaches of digital humanities, text enhances code making it more visible and intelligible for the humanities community."[79]

Zundert, Smiljana Antonijević, and Andrews stress the need for making codework both visible and reputable within the humanities:

> A strategy for making code and codework visible, understandable, and reputable within humanities scholarship must be comprehensive, both in the sense of accounting for the source code and the executed result of software, and by including all relevant stakeholders.[80]

They go on to further stress the belief that:

> theoretical discussions of codework should become an established trajectory in the humanities, along with the development of methods for documenting, analyzing, and evaluating code and codework.[81]

Monfort advocates for the inclusion and incorporation of computer programming practices for the arts and humanities by arguing that we should "understand computation as part of culture."[82] Elsewhere, van Zundert, Antonijević, and Andrews note that "While code and codework increasingly shape research in all fields of the humanities, they are rarely part of disciplinary discussions, remaining invisible and unknown to most scholars.[83] Furthermore,

[76] Joris J. van Zundert, Smiljana Antonijević, Tara L. Andrews, "Black Boxes and True Color—A Rhetoric of Scholarly Code," draft chapter, (unsure how to footnote this), 21.

[77] Bernhard Rieder, Theo Röhle, "Digital Methods: From Challenges to Bildung" Mirko Tobias Schäfer and Karin van Es, *The Datafied Society. Studying Culture through Data*, 115.

[78] Joris J. van Zundert, Smiljana Antonijević, Tara L. Andrews, "Black Boxes and True Color—A Rhetoric of Scholarly Code," draft chapter, (unsure how to footnote this), 23.

[79] Joris J. van Zundert, Smiljana Antonijević, Tara L. Andrews, "Black Boxes and True Color—A Rhetoric of Scholarly Code," draft chapter, (unsure how to footnote this), 21.

[80] Joris J. van Zundert, Smiljana Antonijević, Tara L. Andrews, "Black Boxes and True Color—A Rhetoric of Scholarly Code," draft chapter, (unsure how to footnote this), 20.

[81] Joris J. van Zundert, Smiljana Antonijević, Tara L. Andrews, "Black Boxes and True Color—A Rhetoric of Scholarly Code," draft chapter, (unsure how to footnote this), 21.

[82] Nick Montfort, Exploratory Programming for the Arts and Humanities (The MIT Press, 2016), 1.

[83] Joris J. van Zundert, Smiljana Antonijević, Tara L. Andrews, "Black Boxes and True Color—A Rhetoric of Scholarly Code," draft chapter, (unsure how to footnote this), 19.

they argue that "codework is necessarily shaped by its social context, which may positively or negatively influence the attitude and perception that both coders and other scholars hold towards their work."[84] These are innovative and refreshing perspectives, because they paint coding and computer software in a manner that differs markedly to how it is portrayed not only among the humanities, but among DH researchers, as objective, neutral, or unimportant. "Code can be regarded as a performative application or explanation of theory, rather than a written account of it."[85] Code need not be seen as antithetical to the requirements of big data for the humanities, rather code should be understood as performative, as being in possession of an aesthetics, of having the capacity to display a poetics of code:

> code and codework are all too often treated as an invisible hand, influencing humanities research in ways that are neither transparent nor accounted for. The software used in research is treated as a black box in the sense of information science—that is, it is expected to produce a certain output given a certain input—but at the same time it is often mistrusted precisely for this lack of transparency. It is also often perceived as a mathematical and thus value neutral and socially inert instrument; moreover, these two seemingly contradictory perceptions need not be mutually exclusive.[86]

Concordant with this recognition of the importance and usefulness, indeed the centrality of code as a means of realising the aspirations of big data within the humanities are the heretofore largely unexamined epistemological and methodological issuer that arise from coding in and of itself, and specifically in relation to coding in a humanities environment.[87]

This has been tangentially addressed by numerous scholars via the metaphor of the "black box." Johannes Paßmann and Asher Boersma argue for transparency in our approach to black boxes, making in the process the observation (apparent elsewhere throughout this paper in relaiotn to data) that different authors interpret a term like "transparency" very differently:

> What we found is that different authors have given significantly different meanings to the term transparency. We would like to differentiate between two notions here. On the one hand, there is what we call formalized transparency, which largely tries to obtain 'more positive knowledge' on 'the content' of a black box. On the other hand, we see practical transparency, which does not try to open black boxes, but to develop skills without raising the issue of openability. These two concepts of

[84] Joris J. van Zundert, Smiljana Antonijević, Tara L. Andrews, "Black Boxes and True Color—A Rhetoric of Scholarly Code," draft chapter, (unsure how to footnote this), 23.

[85] Joris J. van Zundert, Smiljana Antonijević, Tara L. Andrews, "Black Boxes and True Color—A Rhetoric of Scholarly Code," draft chapter, (unsure how to footnote this), 16.

[86] Joris J. van Zundert, Smiljana Antonijević, Tara L. Andrews, "Black Boxes and True Color—A Rhetoric of Scholarly Code," draft chapter, (unsure how to footnote this), 1.

[87] Joris J. van Zundert, Smiljana Antonijević, Tara L. Andrews, "Black Boxes and True Color—A Rhetoric of Scholarly Code," draft chapter, (unsure how to footnote this), 21.

> transparency do not exclude each other. Rather, they outline two different sets of practices dealing with black boxes which can complement each other.[88]

Mirko Tobias Schäfer and Karin van Es argue for a transparency that outlines the skills necessary for researchers to deal with the parts of the box that remain “black” or opaque.[89] Similarly van Es, López Coombs and Boeschoten argue the need for

> researchers [to] take responsibility to discern how the given tools work with and shape the data. To fully adopt such a reflexive approach, researchers must consider important questions that relate to each of the three stages of digital data analysis: acquiring, cleaning and analysing.[90]

Again this approach involves openness (transparency) regarding the acquisition and cleaning processes of the data.

With increase in the amounts of data available comes interrogations over whether there are nascent methodological approaches to interacting with this material that we are not happening upon or not thinking up because of an outdated reliance on and indebtedness to the analogue datafication processes. Lev Manovich: "How can our new abilities to store vast amounts of data, to automatically classify, index, link, search, and instantly retrieve it, lead to new kinds of narratives?"[91] Presner gestures towards rethinking the database as a genre: "I wonder how we might rethink the very genre of the database as a representational form”[92] specifically regarding representations of “the indeterminate […] Such a notion of the archive specifically disavows the finality of interpretation, relished in ambiguity, and constantly situates and resituates knowledge through varying perspectives, indeterminacy, and differential ontologies."[93]

Presner explores alternative approaches in the form of distant reading/ distant listening:

> The computational allows us to perform [...] 'distant reading'– a practice that moves away from the close, hermeneutical reading of texts in favor of an algorithmic approach that presents over-arching structures and patterns.[94]

This is a viable approach to the problem of scale in that the scale of events are too large to be realised through conventional narrative mechanisms and one that is similar in its reasoning to Franco Moretti’s association of distant reading with world literature[95] because "the scale, scope, and depth of modernist events doesn't reflect or cannot be captures by the

---

[88] Johannes Paßmann and Asher Boersma, “Unknowing Algorithms: On Transparency of Unopenable Black Boxes” Mirko Tobias Schäfer and Karin van Es, *The Datafied Society. Studying Culture through Data*, 140.

[89] Mirko Tobias Schäfer and Karin van Es, *The Datafied Society. Studying Culture through Data*.

[90] Karin van Es, Nicolàs López Coombs & Thomas Boeschoten, “Towards a Reflexive Digital Data Analysis” ibid., 174.

[91] Lev Manovich, The Language of New Media, 2002, 237.

[92] Presner, in Fogu, Claudio, Kansteiner, Wulf, and Presner, Todd, Probing the Ethics of Holocaust Culture.

[93] Presner, in ibid.

[94] Presner, in ibid.

[95] Franco Moretti, *Distant Reading* (Verso Books, 2013).

structures of storytelling in a realistic mode of narration."[96] A computerised approach then, and particularly one that makes use of visualisations can potentially be an antidote to that; as they allow us to transcend the limitations of linear narrative.

> the structures and meaning-making strategies reserved for historical realism, which was part and parcel of the tradition of storytelling with clear agents, a coherent plot, and narrative strategies characterized by the unities of time, place, and action that gave rise to the logic of a story. In other words, in modernism we see a breakdown of the homology between real events (*Geschichte*) and the narrative strategies (*Historie*) used to represent, capture, communicate, and render these events meaningful.[97]

The complexity of historical events, which are themselves overly simplified when confined to a linear text-bound narrative, could potentially find a suitable medium through which they could be relayed/ represented. And for the conception of this complexity, its scale, its multifaceted-ness; in short, its cultural *richness*, the computational has a lot to offer that traditional narrative modes cannot. In a way then the computational could transcend the challenges to historical representation implicit to analogue narrative models. Visualisation could provide access to and comprehension of scale that would otherwise not be possible: "In other words, the potential to facilitate an ever deeper relationality among the data in a database is one of the conditions of possibility for an ethics of the algorithm.[98]

Writing on algorithms and code etc. is very much presented by those that are familiar and fluent in this process[99] as a creative process and as one that can and should be considered as narrative, just not one humanities (or even DH) scholars are attuned to recognise. How does this change things for us? Is it possible, as Presner envisions, to have "an information architecture that is fundamentally connected to the content […] fundamentally connect[ing] testimony to the information architecture, the data ontologies, the data structures, the indexing systems, and the viewers who are engaged in a participatory mode of listening."[100] Presner:

> we might imagine how a fluid data ontology might work, by allowing, multiple thesauruses that recognize a range of knowledge, standards, and listening practices. For example, what if verbs that connected action and agent, experience and context were given more weight than hierarchies of nouns primarily in associative relationships? What if more participatory architecture allowed for other listeners to create tags that could unsay the said, or in other words, undo —or, at least, supplement—the definitive indexing categories and keywords associated with the segmented testimonies? Or more radically, what if the user interface was generated by network graphs or visualizations, such that the listener did not merely type terms

[96] Presner, in Fogu, Claudio, Kansteiner, Wulf, and Presner, Todd, *Probing the Ethics of Holocaust Culture*.

[97] Presner, in ibid.

[98] Presner, in ibid.

[99] Joris J. van Zundert, Smiljana Antonijević, Tara L. Andrews, "Black Boxes and True Color—A Rhetoric of Scholarly Code," draft chapter, (unsure how to footnote this), 14.

[100] Presner, in Fogu, Claudio, Kansteiner, Wulf, and Presner, Todd, Probing the Ethics of Holocaust Culture.

> into an empty search box but rather could browse the entirety of the archive in a dynamic way based on, perhaps, communities of experience, narrative structure, or even silences, gaps, and so-called non-indexical content?[101]

But is such a fluid user-responsive interface also open to/ vulnerable to corruption?
And this is presuming that the user will be without human bias, which is impossible.
Because human bias will in and of itself create further hidden data, reacting as it would to trends and variables in research interests etc. Are we looking for a database that is truly post-human then? In that it is not only fundamentally without bias, but it evades the possibility of there being human interference, human hierarchising, human bias? This in itself is a bias of course, the imposition of no bias. So rather than there being an ethics of the algorithm, we would have an ethics imposing database, a database/ algorithm that performs ethical unbias, a de-biasing (de-humanising) machine.

2.1.5 Information architecture and the representation of what data are.

Presner's breakdown of the development of the VHA, or as he puts it "the digitization of the Holocaust archive and its transformation into an information management system" [102] is of interest because it provides an insight into the structures that facilitate a digital database:

> This information architecture consists of several components: First, there is the interface itself, which runs in a web-browser, allowing a user to type in keywords, names, and other search terms in order to listen to segments of testimony [...]; behind that, is a relational and structured query language database [...] in which content is organized into tables, records, and fields [...]; all of this data was inputted after the videos themselves were indexed with keywords and other associated information was manually entered [...]. But before this indexing could happen, standards and protocols—which were derived from the National Information Standards Organization's Z39.19 standard for the construction, format, and management of monolingual controlled vocabularies—provided the guidelines for what and how to index the content of the videos. The standard governed the creation of a unique thesaurus to achieve consistency in the description of the content through a controlled vocabulary and thereby facilitate its search and retrieval.[103]

The user interface is replete with a search function that facilitates access to the contents of the server storage via the language database which consists of manually entered keywords along with other metadata that facilitates the users' navigation through the content of the archive. In addition, our conceptions of what data are is influenced by the prevalence of the following "aides" to scholarly research and archives; many of these are remnants of analogue documentation practices that have been carried over to a digital environment: Abstracts, finding aids/ the search function, indexes, keywords, keyword visualisations, metadata, metadata structures. Many of these can be considered prominent examples of

[101] Presner, in ibid.

[102] Presner, in Fogu, Claudio, Kansteiner, Wulf, and Presner, Todd, *Probing the Ethics of Holocaust Culture*.

[103] Presner, in ibid.

skeuomorphism. Digital skeuomorphism are graphical user interfaces that intimates the aesthetics of physical objects.[104] The merits of skeuomorphism versus alternative interfaces have yet to be fully explored.

1. The abstract:

> In the task of 'collectivizing' knowledge, which is truly of our time, the *documentary analysis* or 'abstract' has appeared as one of the most rapid and most reliable means of announcing and communicating thought. [...] Data processing responds to the needs of a research that works upon masses of documents with easily codified statistical indices.[105]

But this in and of itself is problematic because "for unknown texts, description is argument; description will elicit the variety of their discourses."[106] Furthermore, an abstract is narrative, and so a shortened more concentrated abstraction of the main narrative, a proxy narrative. The collation of "related" materials into a data set is problematic in terms of the decisions made regarding inclusion/ exclusion: "a data set is already interpreted by the fact that it is a set: some elements are privileged by inclusion, while others are denied relevance through exclusion."[107]

2. Finding Aids/ the search function

The prevalence of the finding aid or search function in digital user interfaces is a further example of crossover from analogue information science: "Finding aids, which may be available in print or online, provide hierarchical access to collections, boxes of materials, and individual items."[108] However, what Gruber Garvey refers to as "the power of search itself"[109] is and remains under-examined despite the fact that knowledge organisation frameworks are not neutral. Rosenberg similarly argues that: "as humanists, we need to pay much better attention to the epistemological implications of *search*, an entirely new and already dominant form of inquiry, a form with its own rules, and with its own notable blind spots both in design and use."[110]

One of these blind spots comes in the form of the ethical implications behind decisions regarding what is considered worthy of tagging for the purpose of a search and indeed whether this tagging is carried out manually or whether it is machine based. The assignation of a search function sets in place a hierarchizing process from the onset, privileging the

[104] Joris J. van Zundert, Smiljana Antonijević, Tara L. Andrews, "Black Boxes and True Color—A Rhetoric of Scholarly Code," draft chapter, (unsure how to footnote this), 15.

[105] Ibid., 15, emphasis in original.

[106] Williams, "Procrusteam Marxism and Subjective Rigor: Early Modern Arithmetic and Its Readers," in," in Gitelman, *"Raw Data" Is an Oxymoron*, 41.

[107] Williams, "Procrusteam Marxism and Subjective Rigor: Early Modern Arithmetic and Its Readers," in," in ibid.

[108] Borgman, "Big Data, Little Data, No Data," 162.

[109] Ellen Gruber Garvey, "'facts and FACTS:' Abolitionists' Database Innovations," Gitelman, *"Raw Data" Is an Oxymoron*, 91.

[110] Rosenberg, "Data before the Fact," in ibid., 35, emphasis in original.

facets summarised in the metadata (facets that are more readily searchable), facets that have been highlighted by the documentator or archive curator as worthy of interest, and so on.

What's important here is that data's status as rhetorical and performative is acknowledged and not impeded or overlooked so that in the context of specific digitized archives (etc.) this status is not manipulated in order for it to be rhetorical or performative in specific or restricted ways, or according to predetermined rhetorical tropes/ desired performativity. A (perhaps controversial) example is the SHOAH VHA, which provides a massive quantity of data, but this data has limited rhetorical potent, or its capacity has been delimited. Say, for example, one wanted to research racism amongst Holocaust survivors, the keyword database would be of no use to you; the performative contract of the material has been predetermined and delimited by the architects of the project. This in itself is understandable, as the SHOAH VHA was conceived to deliver a set narrative and to perform a specific series of functions. In other words, unforeseen delimitations can be brought about by software functions.

3. Keyword indexes and visualisation; the SHOAH Visual Archive:

In the case of the SHOAH VHA we see the adoption of standardised vocabulary through their adherence to the NISO Z39.19. NISOZ39.19's goal is: "to provide 'guidelines for the selection, formulation, organization, and display of terms that together make up a controlled vocabulary' for the purpose of 'knowledge management' and 'knowledge organization.'"[111] The keyword principles "derive from the application of a specific standard (Z39.19) to consistently and unambiguously describe 'content objects' (the survivor testimonies) in order to produce a monolingual controlled vocabulary (the thesaurus) to facilitate their search and retrieval."[112] Hierarchical vocabularies/ taxonomies form the structure of search programmes such as the SHOAH VHA, and these are tiered so that "under each of these broad categories are hierarchical vocabularies to facilitate searching at a more precise level."[113]

Presner notes that

> The question of what constitutes a keyword is the starting point for query design, for that is what makes querying and query design practically part of a research strategy. When formulating a query, one often begins with keywords so as to ascertain who is using them, in which contexts and with which spread or distribution over time.[114]

I would argue that the question as to what constitutes a keyword it also the starting point for epistemic and ethical queries. The SHOAH VHA keyword index is problematic because the keyword indexing system was not done automatically, but manually, and moreover because

[111] Fogu, Claudio, Kansteiner, Wulf, and Presner, Todd, *Probing the Ethics of Holocaust Culture*, Quoted in Presner,.

[112] Presner, in Fogu, Claudio, Kansteiner, Wulf, and Presner, Todd, Probing the Ethics of Holocaust Culture.

[113] Ibid.

[114] Richard Rogers, "Foundations of Digital Methods: Query Design" Mirko Tobias Schäfer and Karin van Es, *The Datafied Society. Studying Culture through Data*, 81.

the keyword system "is the *only* way to search the content of the testimonies."[115] Presner argues that with respect to the SHOAH VHA: "we see a symbiosis between narrative and database, such that the paradigmatic structure of the database contributes to the syntagmatic possibilities of combination at the heart of narrative."[116] Subjective selection for the purpose of narrative formation takes place at a structural level within this archive, and the "syntagmatic possibilities" are limited to those imposed by the people who assigned the keywords.

Keyword indexes can also be used to counteract this inclusion/ exclusion search methodology; thus they can be part of "programmes or anti-programmes":

> keywords can be part of programmes or anti-programmes. Programmes refer to efforts made at putting forward and promoting a particular proposal, campaign or project. Conversely, anti-programmes oppose these efforts or projects through keywords. Following this reading, keywords can be thought of as furthering a programme or an anti-programme. There is, however, also a third type of keyword I would like to add, which refers to efforts made at being neutral. These are specific undertakings made not to join a programme or an anti-programme.[117]

Presner classifies the keywords that constitute the body of material in the VHA search system as "meta-data scaffolding" in the form of keywords used as scaffolding for a management system that facilitates keyword searches; thereby implicitly acknowledging the rhetorical function of the search function: "the meta-data scaffolding and data management system, including the numerous patents for its information architecture, that allows users to find and watch testimonies."[118] This is the *only* way to search through the archive in its totality, so this in itself is problematic because it limits searchability to the keywords chosen manually by human project contributors.

Indexes within a digital environment also allow for multiple associations and facilitate a richer network:

> SHOAH VHA allows for more than one association per index element, instead as many as three kinds of relationships can exist between any two (or more) indexing elements. These are "Inheritance, whole/ part, and associative relationships."[119]

Finally, aside from the hierarchy imposed by a keyword facilitated search system, search systems in general are misrepresentative precisely because they merely facilitate the user accessing the totality of the database, not the totality of the events represented by the data contained within the database. The database itself then can be considered a highly fabricated (cooked?) simulacrum of a real life happening, event, or entity; it is not the thing in and of itself, but a recording of the thing, and a recording that is always incomplete, always

[115] Presner, in Fogu, Claudio, Kansteiner, Wulf, and Presner, Todd, *Probing the Ethics of Holocaust Culture*.

[116] Presner, in ibid.

[117] Richard Rogers, "Foundations of Digital Methods: Query Design" Mirko Tobias Schäfer and Karin van Es, *The Datafied Society. Studying Culture through Data*, 82.

[118] Presner, in Fogu, Claudio, Kansteiner, Wulf, and Presner, Todd, *Probing the Ethics of Holocaust Culture*.

[119] Samuel Gustman, quoted in Presner, ibid.

pre-determined and always already subject to interpretative manipulations and machinations, both visible and hidden.

4. Visualisations (visual representations of data)

Rosenthal: "visualizations have long been central to a certain branch of data research known as 'exploratory data analysis.'"[120] Visualisations can take the form of network topologies, timelines or enriched cartographies,[121] among others; and again many of these are merely developments on long-established analogue practices. Rather like the discipline and skill miscibility discernible identifiable in the production of data, discussed previously, visual representations of data, as Drucker observes, are largely predetermined by the disciplines that engendered them:

> The majority of information graphics […] are shaped by the disciplines from which they have sprung: statistics, empirical sciences, and business. Can these languages serve humanistic fields where interpretation, ambiguity, inference, and qualitative judgment take priority over quantitative statements and presentations of "facts"?[122]

It remains to be seen as to whether these tools can be adapted to facilitate research in an environment that thrives on multivalent and tiered levels of uncertainty akin to the "uncertainty matrix" discussed previously. One major nuisance of the visual model of representing data is, as Schäfer and van Es observe, the tendency of users to believe what they see, to take the visualisation as an objective and truthful rendering: "Visualized via colourful dashboards, infographics and charts, it puts forth, persuasively and seductively, a seemingly accurate and unbiased assessment of reality."[123] Schäfer and van Es also echo Rosenthal in their urge for caution apropos the untempered use of visualisations without considered critical reflection:

> Our current enthusiasm for computer-aided methods and data parallels the technology-induced crisis in representation and objectivity analysed by Daston and Galison. Their concerns must be taken into account in order to critically reflect upon the purported objectivity of computer-calculated results and visualizations.[124]

Rosenthal highlights the prominent role of visualisations as a key difference between traditional analogue-driven literary criticism and data-driven literary criticism, adopting a proto-cognitive approach to the study and effect of data visualisation:

> traditional literary criticism in your mind looked rather like this page: words, sentences, paragraphs. Depending on how materialist your critical tastes are, there might have been an image or two, but they likely captured an artefact of the time

[120] Rosenthal, "Introduction," 5.

[121] Bernhard Rieder, Theo Röhle, "Digital Methods: From Challenges to Bildung" Mirko Tobias Schäfer and Karin van Es, *The Datafied Society. Studying Culture through Data*, 112.

[122] Johanna Drucker, *SpecLab* (Chicago: University of Chicago Press, 2010), 6–7, http://www.press.uchicago.edu/ucp/books/book/chicago/S/bo6211945.html.

[123] Mirko Tobias Schäfer and Karin van Es, *The Datafied Society. Studying Culture through Data*, 13.

[124] Ibid., 17–18.

> period under discussion. I would bet that the data-driven literary criticism, on the contrary, contained visualizations: graphs, maps, or trees (to use the title of one of Moretti's [2005] books). And these visualizations are not just there to supplement the argument: they go a long way toward making the argument. This, I would like to suggest, is a principal distinction between the two modes of criticism, and it is one that should demonstrate the connection with narrative and data. On the one hand, we have a retelling — an artful and opinionated reshaping — of the underlying evidence. On the other hand, we have a synchronic, visual representation that attempts to let the data speak for itself, without mediation.[125]

So, traditional narrative driven criticism took the form of reading whereas data driven criticism takes the form of visualisation. This turn to the visual is driven by a hermeneutic belief akin to Gruber Garvey's assertion that "Data will out"[126]:

> [I]n data-driven literary criticism the visualization allows us to engage [...] in a way that language would not. As Edward R. Tufte (2001, 14), a statistician and author on data visualization, puts it, in almost Heidegerrian language, "Graphics *reveal* data."[127]

Critically, however, Rosenthal maintains awareness of data's pre-empirical status, arguing instead that the visual medium facilitates engagement with the data that results in new connections. This is critical, because this new level of engagement is what is proffered by harnessing big data for humanities research:

> Data in this formulation speaks to us (or "seems to" speak to us) through a visual medium. And even more importantly, the information it conveys does not confirm or refute theories we already hold. Instead, it seems to bring something new.[128]

Rosenthal then, represents a faction of scholars that identify the anarchic potential of data and, in particular of visual representations of data: "At the heart of much data-driven literary criticism lies the hope that the voice of data, speaking through its visualizations, can break the hermeneutic circle."[129] This anarchic potential was something that was perhaps muzzled somewhat by the SHOAH VHA's decision to identify facets of the archive perceived as untoward, inappropriate or simply beyond categorisation as "so-called non-indexical content."[130]

Visualisations are often generated based off of keyword indexes:

> computer-generated visualisations [...] based on the general indexing categories developed by the Shoah Foundation to organise the genocide-related concepts and experiences described in the 49,000 Jewish survivor testimonies in the Visual History

[125] Rosenthal, "Introduction," 4.

[126] Ellen Gruber Garvey, "'facts and FACTS:' Abolitionists' Database Innovations," Gitelman, *"Raw Data" Is an Oxymoron*, 90.

[127] Rosenthal, "Introduction," 5.

[128] Ibid.

[129] Ibid., 6.

[130] Presner, in Fogu, Claudio, Kansteiner, Wulf, and Presner, Todd, Probing the Ethics of Holocaust Culture.

> Archive [...]. These categories form the most general or highest level in the 50,000-word thesaurus created by the Foundation.[131]

As we know from the previous analyses of the problems inherent in keyword assignation, the visualisation that accompany these already problematic indexes are far from unbiased; a fact that renders the predisposition of users to "believe" visual representations all the more problematic. Rieder and Röhle speak compellingly on "The Power of Visual Evidence," arguing that "Since these visualizations possess spectacular aesthetic – and thus rhetorical – qualities, we [ask] how the argumentative power of images could (or should) be criticized."[132]

Finally, aside from the hierarchy imposed by a keyword facilitated visual search system, visual search systems and interfaces in general are misrepresentative precisely because they merely facilitate the user accessing the totality of the database, not the totality of the events represented by the data contained within the database:

> Without the visual interface, the database is still searchable by way of the tables containing structured data (name, place of birth, date of birth, date of death, family members, and so forth); however, the totality cannot be seen without an interface that visualizes the scope and scale of the database (which is —and this is critically important—a very different thing than the 'whole' of the event called 'the Holocaust').[133]

Visualisation can be useful because they facilitate "an interrogation of the reality effect produced by such ways of seeing and experiencing."[134] Presner argues that the visual approach is a pragmatic and viable method for facilitating human inquiry into extremely large scale databases: "Visualizations like these might provide new starting points for delving into the more than 6 million records in the database and seeing connections that a human eye could not possibly detect or track."[135]

5. Metadata

What data is is further delimited by what is provided in the metadata and metadata structures of information architecture. The metadata is performatively modifying the performative capacities of the data. Drucker argues that metadata structures have the greatest impact on our approach to material in a digital environ:

> Arguably, few other textual firms will have greater impact on the way we read, receive, search, access, use, and engage with the primary materials of humanities

[131] Todd Presner, "The Ethics of the Algorithm: Close and Distant Listening to the Shoah Foundation Visual History Archive" in ibid.

[132] Bernhard Rieder, Theo Röhle, "Digital Methods: From Challenges to Bildung" Mirko Tobias Schäfer and Karin van Es, *The Datafied Society. Studying Culture through Data*, 112.

[133] Presner, in Fogu, Claudio, Kansteiner, Wulf, and Presner, Todd, *Probing the Ethics of Holocaust Culture*.

[134] Presner, in ibid.

[135] Presner, in ibid., 36.

> studies than the metadata structures that organise and present that knowledge in digital form.[136]

This assertion it granted further credence by Briet's insistence in *What is Documentation* that analogue metadata in the form of indexes and catalogues are the "primary intermediary" between a document and its user; again, this phenomenon has transitioned from analogue to digital collections:

> Current *catalogues*, retrospective catalogues, and union catalogues are obligatory documentary tools, and they are the practical intermediaries between graphical documents and their users. These catalogues of documents are themselves documents of a secondary degree.[137]

Presner similarly observes the pre-eminent influence of the metadata on the SHOAH VHA database, arguing that it operates as a "paratext":

> The metadata database of the Shoah Foundation VHA thus represents a kind of 'paratext' insofar as it can be reordered, disassembled, and reassembled according to the constraints and possibilities of computational logic. The visualization of the Shoah Foundation VHA are representations of the paratext, the metadata scaffolding that runs behind the testimonies and, with every query to the database, represents an algorithmically transformed text.[138]

Again this proves problematic in terms of the crossover of terminology that are not altogether synonymous, with Presner here substituting "metadata" for paratext.

The performative status of metadata can be aligned with Kevin D. Haggerty and Richard V. Ericson's on the "'data double,' our virtual/ informational profiles that circulate in various computers and contexts of practical application."[139] When it comes to "big data" in the humanities, the "data double" cannot supplement the original figure, the object of interest, the figure of research. But as we have seen (Krajewski, 2013; Edmond, 2016), it frequently stands in for the original documentation pertaining the entity. The question remains however over who gets to make the decision over whether the original context is relevant or not.

To make matters even more complicated, in addition to metadata and data being performative, the task of producing data retroactively shapes the contexts that surround it (both in its native state and in the database, so on a multi-tiered level depending on the level of treatment of the data): "the work of producing, preserving, and sharing data reshapes the organizational, technological, and cultural worlds around them."[140] So there is a two-way relationship, with data shaping context, and context shaping data. Thus, metadata's capacity

[136] Johanna Drucker, *SpecLab*, 9.

[137] Briet et al., *What Is Documentation?*, 11.

[138] Presner, in Fogu, Claudio, Kansteiner, Wulf, and Presner, Todd, Probing the Ethics of Holocaust Culture.

[139] Richard Victor Ericson and Kevin D. Haggerty, The New Politics of Surveillance and Visibility (University of Toronto Press, 2006), 4.

[140] David Ribes and Steven J. Jackson, "Data Bite Man: The Work of Sustaining a Long-Term Study" Gitelman, *"Raw Data" Is an Oxymoron*, 147.

to influence the material it curates is underacknowledged, as is code, codework and the algorithms employed to facilitate the digitising process. The growing influence of what Ribes and Jackson refer to as "the invisible infrastructures of data"[141] is similarly acknowledged by Johanna Drucker:

> The majority of information graphics […] are shaped by the disciplines from which they have sprung: statistics, empirical sciences, and business. Can these languages serve humanistic fields where interpretation, ambiguity, inference, and qualitative judgment take priority over quantitative statements and presentations of "facts"?[142]

2.1.6 A summary of why incomplete data streams are so prevalent.

Presner introduces the concept of a "data sublime,"[143] a state or system that can successfully represent "the vastness of the different accounts of the events in question."[144] Is the "data sublime" a simulacrum for the event "as it happened? The term "sublime" suggests an ideal (sublime data), a successful re-presentation or networking of "the vastness of the different accounts of the events in question."[145] But data sublime can also suggest a complete or near perfect form (data that has been fully sublimated). Nevertheless, were it to exist a "data sublime" would be hindered by the software itself because the information architecture that structure it is necessarily literal and unambiguous: "the data sublime [...] is structured by an information management system that is remarkably literalist."[146] The data sublime might intimate or suggest the possibility of completeness, but such completeness is difficult if not impossible to achieve for the following reasons:

1. Because "all that data would be meaningless without an organizing scheme."[147] So, because metadata standards and the mediating effect of metadata on that which is accessible in the archive.

2. In the case of the humanities, because of specificities regarding the scale and complexity of (historical) events in and of themselves. Presner: "The first concerns the scale, scope, and complexity of the events themselves."[148]

3. Because data is by nature an incomplete and partial snapshot of the thing in and of itself; it is not a substitute for "totality," at least not yet. Presner refers to this as the "lack of homology between the reality of 'what happened' and the modalities of

[141] David Ribes and Steven J. Jackson, "Data Bite Man: The Work of Sustaining a Long-Term Study" ibid., 152.

[142] Johanna Drucker and Bethany Nowviskie, "Speculative Computing: Aesthetic Provocations in Humanities Computing," in *A Companion to Digital Humanities* (Blackwell Publishing Ltd, 2007), 6–7, http://onlinelibrary.wiley.com/doi/10.1002/9780470999875.ch29/summary.

[143] Presner, in Fogu, Claudio, Kansteiner, Wulf, and Presner, Todd, *Probing the Ethics of Holocaust Culture*.

[144] Presner, in ibid.

[145] Presner, in ibid.

[146] Presner, in ibid.

[147] William Uricchio, "Data, Culture and the Ambivalence of Algorithms," Mirko Tobias Schäfer and Karin van Es, *The Datafied Society. Studying Culture through Data*, 125.

[148] Presner, in Fogu, Claudio, Kansteiner, Wulf, and Presner, Todd, *Probing the Ethics of Holocaust Culture*.

representation, whether through narrative, visual, or computational techniques."[149]

4. Because the act of perceiving is always subjective and partial. Presner notes "the problem of limited human facilities to observe, comprehend, read, listen to, and finally adjudicate the vastness of the different accounts of the events in question."[150]

5. Because information management systems are literalist. From what we looked at in the section that addressed code, this is not necessarily something that has to continue; if integrated properly, it could be possible to work against this delimiting factor.

6. Because of the metadata structures and search functions that governs accessibility to the material. Functions such as: Keyword assignation, which is not always an ongoing process and does not always allow for input from users. This is the case in the SHOAH VHA:

   Keywords were assigned to the narrative content of the video from the thesaurus and, at the same time, new keywords could be proposed to describe experiences not already in the thesaurus. [...] Not every minute segment, however, has a keyword, something that may indicate the continuation of the previous keyword but may also mean, [...] 'lack of indexable content.'[151]

Incomplete data streams are also so prevalent because of two actions, one active and one passive, that influence the accessibility of the archive. First, the issue regarding how to approach unindexable content; what Presner refers to as "lack of indexable content":

> Lack of indexable content can mean many things, ranging from an interviewer asking a question to a survivor repeating him or herself, a pause in the conversation to reflect search for the right words, an emotional moment, noise, silence, or even content that the indexer doesn't want to draw attention to (such as racist sentiments against Hispanics, for example, in one testimony). In other words, indexable content is manifest content in a declarative or imperative mode —in general, what is literally and objectively said. Altogether, the indexing system produces a kind of 'normative story' (purged of certain contingencies and unwanted elements) in which — on the level of the data in the database—many of the testimonies, but certainly not all, become quite like each other.[152]

Second, the sections of the dialogues (comprising the vast majority of the archive of testimonies) that are *not* assigned a keyword. An analogy can be made to John Cage's concept of the dynamic interdependency between sound and silence. Sound is musical notation, but sound relies on being surrounded by silence, and vice versa. The Cage analogy is particularly apt given the popularity of phrases such as "separating signal from noise," when speaking about data cleaning or processing. Privileging one neglects and

[149] Presner, in ibid.

[150] Presner, in ibid.

[151] Presner, in ibid.

[152] Presner, in ibid.

undercuts the importance of the other. Each section assigned a keyword has a relationship with the material on either side of it within the dialogue, but this material is left "silent," "hidden" (in certain cases, especially cases such as the one above where certain facets of the testimonies are "purged" (Presner's word) from the keyword thesaurus of the archive), or alternatively left latent or unstructured in the archive.

Again, Presner talks about expelling the "latent content" but there's additional latent or unstructured content to the content he lists below:

> The result is a massive data ontology that has expelled the latent content, the performative, the figural, the subjunctive, the tone of questioning and doubt, the expressiveness of the face and the very acts of telling (and failing to tell) that mark the contingency of all communication. And while its aim is objectivity, it is important to underscore that a human listener decided what to index and what not to index; a human listener decided what indexing term to use and that indexing term not to use; and a human listener decided if a given narrative segment could be described by a keyword or not.[153]

Presner justifies the SHOAH approach by arguing that it is a larger scale version of "what historians" do:

> insofar as they employ events at various levels of 'zoom' in order to convey different kinds of meaning. In other words, we 'toggle' back-and-forth between macro-level accounts of the totality of the event (zoomed out) and the micro-level accounts of individual experiences (zoomed in), which are, by their very nature, defined by specific experiences, perspectives, spectatorship, language, and so forth[154]

The justification that "this [the SHOAH VHA setup] is not very different from what historians do"[155] is not sufficient; because it neglects to take into account the prescriptive effect this structure has on future scholars and users. A singular small scale study by a historian is one thing, we can accept that there may be biases, and we are sensitive/ alert to the fact that the materials here have been structured and selected in such a way as to present a specific argument. But in the case of a larger database, the same argument becomes redundant, and potentially dangerous.

> 1. The influence of unambiguous and ambiguous queries, and the difficulty of quantifying and standardising what exactly qualifies as "unambiguous" and "ambiguous" in relation to humanities research:

> If you peruse the search engine literature, there are mentions of navigational queries, transactional queries and substantive queries, among other types. Yet, on a meta-level, we can broadly speak of two kinds of queries: unambiguous and ambiguous.[156]

[153] Presner, in ibid.

[154] Presner, in ibid.

[155] Presner, in ibid.

[156] Richard Rogers, "Foundations of Digital Methods: Query Design" Mirko Tobias Schäfer and Karin van Es, *The Datafied Society. Studying Culture through Data*, 87.

2. Because of the misconception that data streams should be/ can be complete in the first place; that they represent totality.

3. Because the methodologies adopted by computer scientists create models that are, and can only ever be, incomplete:

> A majority of [computer science] papers today also use supervised machine learning, an automatic creation of models that can classify or predict the values of the new data using already existing examples. In both cases, a model can only account for part of the data, and this is typical of the statistical approach.[157]

In conclusion, there is a need to radically reassess what exactly we want our databases to do. van Zundert, Antonijević, and Andrews stress the fact that "A practical examination and theoretical discussion of how software reflexively interacts with humanities research remains an urgent necessity."[158] Do we want our digital research environments to mimic, or continue to mimic analogue methodology in the form of digital skeumorphisms that are themselves flawed and prone to error and human bias? Is it possible to design something that goes beyond bias? In addition, one of the reasons given in section 2.1 is that many databases are organised and managed by systems that are "remarkably literalist"[159] There appears to be little awareness among computer scientists of the implications of this literalism? There is a clear need to explore other forms of Knowledge Representation Schemes and alternatives to metadata, information management systems, organisational structures and search functions.

2.2. Data defined.

An overview of the etymology of the term data and the history of its usage is provided by Rosenberg[160] while Christine Borgman in *Big Data Little Data No Data* provides a thorough run-through of various definitions of data, invaluable for the compilation of this overview and for outlining the taxonomic models used, thanks to her inclusion of detailed case-studies from the disciplines of Sciences, Social Sciences, and the Humanities.

Rosenberg also outlines the early history of "data" as a concept prior to the 20th century, exploring how it acquired its "pre-analytical, pre-factual status."[161] Rosenthal similarly presents data as an entity that resists analysis:

> The term, at least as we use it now, refers to an object that resists analysis. [...] this assertion of data's pre-epistemic irreducibility has a certain amount of tautology. Data

[157] Lev Manovich, "Cultural Analytics, Social Computing and Digital Humanities," ibid., 64.

[158] Joris J. van Zundert, Smiljana Antonijević, Tara L. Andrews, "Black Boxes and True Color—A Rhetoric of Scholarly Code," draft chapter, (unsure how to footnote this), 4.

[159] Presner, in Fogu, Claudio, Kansteiner, Wulf, and Presner, Todd, *Probing the Ethics of Holocaust Culture*.

[160] Daniel Rosenberg, "Data before the Fact,"" Lisa Gitelman, ed., *"Raw Data" Is an Oxymoron*, Infrastructures Series (Cambridge, Massachusetts ; London, England: The MIT Press, 2013), 15–40.

[161] Rosenberg, "Data before the Fact," in Gitelman, "Raw Data" is an Oxymoron, 18.

> is different from facts and evidence not because it is some purer object but because it is defined precisely as that object that cannot be questioned.[162]

Borgman elaborates, stating that "Data are neither truth nor reality. They may be facts, sources of evidence, or principles of an argument that are used to assert truth or reality."[163] Rosenberg cites data as rhetorical: “facts are ontological, evidence is epistemological, data is rhetorical.”[164] He also reminds us that “False data is data nonetheless”[165] and that “Data has no truth. Even today, when we speak about data, we make no assumptions about veracity.”[166] That last statement is important, because while, when we speak of data we perhaps make no assumptions about truth, when we speak about big data, we do, as the 5 characteristics of big data being “volume, velocity, variety, veracity and value.” Somewhere in the transition from data to big data, veracity becomes important; even if the datasets that compose the big data are themselves, when taken on a small scale, representative of data as having “no truth,” as “pre-factual” or any of the conceptual interpretations of data we have just encountered. These conflicting conceptual definitions pose a number of problems: How can data be pre-epistemic but also represent, as Borgman argues they do, “forms of information”? How can data be “facts” if they are pre-factual? How can (small) data bring no assumptions of veracity, but big data be veracious?

Speaking in the context of data’s relationship to fictional literary narratives, Rosenthal identifies data as a “fiction”: “The fiction of data, the illusion of data."[167] In addition, Rita Raley posits data, and in particular the “data-double” to a given entity, as “performative”: “Our data bodies then are repeatedly enacted as a consequence of search procedures. Data is in this respect performative.”[168] Raley also inadvertently highlights a problematic facet of discourse on data: the language used is overly theoretical and alienating, this is problematic given the cross-disciplinary nature of “data” and for those approaching these debates from a computer science background. It does not encourage cross-disciplinary dialogues. David Ribes and Steven Jackson contribute to Raley’s notion of data as performative by affixing a further performative capacity to data, the capacity for data to affect its contexts: “In this context [of a long-term study], data—long-term, comparable, and interoperable—become a sort of actor, shaping and reshaping the social worlds around them.”[169] This is analogous to Floridi’s “ethics of information.” Ribes and Jackson discuss examples of “ecologies, scientific objects, and data archives that exemplify the ways phenomena shape the social order that seek produce, manage, and preserve them.”[170] So, data can be rhetorical, performative, and fictional, but also pre-factual, pre-epistemic, and while “neither truth nor reality […] used to

[162] Jesse Rosenthal, “Introduction: ‘Narrative against Data,’” *Genre* 50, no. 1 (April 1, 2017): 1., doi:10.1215/00166928-3761312.

[163] Christine L. Borgman, “Big Data, Little Data, No Data,” MIT Press, 17, accessed April 7, 2017, https://mitpress.mit.edu/big-data-little-data-no-data.

[164] Rosenberg, “Data before the Fact,” in Gitelman, *“Raw Data” Is an Oxymoron*, 18.

[165] Rosenberg, “Data before the Fact,” in ibid.

[166] Rosenberg, “Data before the Fact,” in ibid., 37.

[167] Rosenthal, “Introduction,” 3–4.

[168] Rita Raley, “Dataveillance and Countervailance” Gitelman, *“Raw Data” Is an Oxymoron*, 128.

[169] David Ribes and Steven J. Jackson, “Data Bite Man: The Work of Sustaining a Long-Term Study” ibid., 148.

[170] David Ribes and Steven J. Jackson, “Data Bite Man: The Work of Sustaining a Long-Term Study” ibid.

assert truth or reality"[171] while also being capable of falsity, because "False data is data nonetheless,"[172] and possessing a capacity to reconfigure its environs and interlocutors. Drucker (2011) and Kitchin (2014) make the distinction between data ('given' in Latin) and capta ('taken'), each preferring the latter term[173]:

> Drucker, speaking more specifically of information visualizations, takes this a step further, arguing that the very data we use are already infused with interpretation. Rendering information in graphical form, she claims, 'gives it a simplicity and legibility that hides every aspect of the original interpretative framework on which the […] data were constructed' (2014: 128). Drucker's point here is that data are always preconstituted, shaped by the parameters for their selection. Others have stressed that these parameters are never neutral, but construct the world as profoundly ideological (e.g. Posner 2015).14 Therefore, we are well-advised to think of them not as data (given) but rather as capta (taken), 'constructed as an interpretation of the phenomenal world' rather than inherent to it (Drucker 2014: 128).[174]

Irrespective of whether it's referred to as data or capta, however, that which constitutes data (or capta) are culturally specific, idiosyncratic, and variable. William Uricchio observes that "the rules for what constitute data, together with the logics of their assembly, make up a core component of culture. Whether they be omens or numbers, whether they are qualitative or quantitative, whether they involve heuristics, hermeneutics or the rules of mathematics, the dyad of data and their organizing schemes give cultural eras their specificity."[175] What constitutes data is often idiosyncratic and unclear. Data, it seems, can be anything. Jesse Rosenthal observes that "When we call something data, we define what we can do with it and what we can say about it."[176] The problem is that what it is, what can be done with it, and what can be said about it—and conversely our understandings of what data itself can do to us, to its interlocutors, environments and to its contexts—vary drastically from discipline to discipline, and even within disciplines. In addition, this overabundance of rhetorical strategies regarding what data is and how we are to speak about it also serves to create a certain distance from data, granting data a certain objectivity that belies its curated, malleable, reactive, and performative nature.

To make things even more complex, any definition of data (or the architecture that makes it available in an analogue or digital environment) needs to maintain an awareness of the speculative potential of the information contained within its datasets. Remarking on definitions of data offered by Peter Fox and Ray Harris (2013) and Paul Uhlir and Daniel Cohen (2011)—definitions that work by example or by assigning attributes to examples of data— Borgman observes "any such list is at best a starting point for what could be data to

---

[171] Borgman, "Big Data, Little Data, No Data," 17.

[172] Rosenberg, "Data before the Fact," in Gitelman, *"Raw Data" Is an Oxymoron*, 18.

[173] Karin van Es, Nicolàs López Coombs & Thomas Boeschoten, "Towards a Reflexive Digital Data Analysis" Mirko Tobias Schäfer and Karin van Es, eds., *The Datafied Society. Studying Culture through Data* (Amsterdam University Press, 2017), 173, http://www.oapen.org/search?identifier=624771.

[174] Eef Masson, "Humanistic Data Research An Encounter between Epistemic Traditions," in ibid., 32.

[175] William Uricchio, "Data, Culture and the Ambivalence of Algorithms," ibid., 125.

[176] Rosenthal, "Introduction," 2.

someone, for some purpose, at some point in time."[177] Not only can data be anything extant then, it is conjectural, notional, speculative, which again adds credence to Raley's notion of data as performative. As a performative entity it has the potential to be *other*; to assume, perform, and acquire other values. So data has the potential to be subversive, and subversive in a manner akin to the performative as it is forwarded throughout the writings of Judith Butler.

Inconsistent and contradictory statements on data are frequent; even within a collection of critical essays held under the unificatory title *Raw Data is an Oxymoron*. For example, Ellen Gruber Garvey asserts that "Data will out."[178] This implies the opposite to Rosenberg's statement that data has a "pre-analytical, pre-factual status"[179] and Borgman's similar observation that "Data are neither truth nor reality. They may be facts, sources of evidence, or principles of an argument that are used to assert truth or reality."[180] Raley picks up on this contradiction when she outlines the arguments for and against data surveillance, noting that such back and forth arguments "imply that data is somehow neutral and that it is only the uses of data that are either repressive or emancipatory"[181]; and indeed data was used to great emancipatory effect throughout *American Slavery As It Is.* However, Gruber Garvey's assertion that "data will out" implies that there is some overarching "truth-function" to the data, that it can somehow speak for itself, as opposed to needing context and modification so as to take on the values expected/ sought out by the researcher. The data only "outs" by means of a complex set of procedures that may be discipline specific, case specific, or even researcher specific. These procedures are also idiosyncratic, often conceived of on the fly, and so concordantly they display methodological approaches *in progress* that are left undocumented and unaccounted for. In the case of Gruber Garvey's essay on *American Slavery As It Is*, to cite but one example, "the ads were abstracted, their information pried loose and accumulated, aggregated en masse."[182] This data did not simply "out," it was forcibly pried out.

Further still, if we are to accept Borgman's observation regarding the speculative potential for any item to be identified as and function as data "to someone, for some purpose, at some point in time"[183] then the materials that makeup or constitute the data will not necessarily always "out" then because the so-called "raw data" we are provided with has undergone extensive, often undocumented cleaning to get it into a state where it is recognisable as data in the first place. Data cleaning in and of itself then is an interpretative act, because in doing so one establishes and distinguishes between signal and noise, and removing or muting the noise so as to highlight the signal. This means that, for example, in the case of *American Slavery As It Is*, the material scrubbed from the data when it was in its native environment (newspapers, etc.) is no longer accessible, having been deemed external to the remit of data selected for inclusion. This expunges items of potential future (speculative) value or merit *as*

[177] Borgman, "Big Data, Little Data, No Data," 19.

[178] Ellen Gruber Garvey, "'facts and FACTS:' Abolitionists' Database Innovations," Gitelman, *"Raw Data" Is an Oxymoron*, 90.

[179] Daniel Rosenbery, "Data before the Fact,"" ibid., 18.

[180] Borgman, "Big Data, Little Data, No Data," 17.

[181] Rita Raley, "Dataveillance and Countervailance" Gitelman, *"Raw Data" Is an Oxymoron*, 130.

[182] Ellen Gruber Garvey, "'facts and FACTS:' Abolitionists' Database Innovations," ibid., 91.

[183] Borgman, "Big Data, Little Data, No Data," 19.

data. That which is not presently identified as data can or could be identified at any given point as data; non-data is thus arguably always potentially speculative data. If anything is potentially data, then anything that is cleaned or scrubbed during the creation of datasets that consist of so-called “raw data” is expunged from the database, and so from this perspective all data streams are, to a certain degree, incomplete, and necessarily so.

It’s hardly surprising then, when faced with these definitions of data that Borgman notes "Data are most often defined by example, such as facts, numbers, letters, and symbols [...] Lists of examples are not truly definitions because they do not establish clear boundaries between what is and is not included in a concept."[184] Indeed it is Borgman, who is perhaps the most concise and lucid writer on this topic, who provides perhaps the most pragmatic observational definition on data:

> *data* has yet to acquire a consensus definition. It is not a pure concept nor are data natural objects with an essence of their own. The most inclusive summary is to say that data are representations of observations, objects, or other entities used as evidence of phenomena for the purpose of research or scholarship.[185]

Despite these variants in how data is observed and what observations are deemed data-worthy, it’s ostensibly clear at least from the title of *Raw Data is an Oxymoron* that there is a round consensus that “raw data is not so raw,”[186] that “raw data” is a misnomer. But the degrees of “rawness” again are variable, discipline-specific, project-specific, and researcher-specific. So-called raw data is subject to extensive cleaning and modification (dealt with later) that is not always acknowledged, fully accounted for, or consistent. In addition, in certain cases, “Only a small portion of what comes to be considered raw data is actually generated in the field.”[187] Ribes and Jackson are illuminating on the processes native data undergoes to achieve “raw status” and, further still, the work that goes into preserving data, observing that “we often think of raw data as following straight and commonsensical pathways from collection to database. Sometimes this is true […] however the more common story [...] [sees] data moving through complex, multi-institutional networks.”[188]

A key facet of this problem is that the term data is ubiquitous, but is consistently interpreted differently, used in different contexts, or to refer to different things, leading to confusion and disorganisation that if left unaddressed, will have significant impact on future research programmes and research infrastructures. There needs to be consensus in terms of what we speak about when we speak about data, irrespective of how difficult it is to “define” it when it comes to the diverse cultural resources that fuel humanities research. I think rather than attempt something akin to a computerised “definition of culture” we would perhaps be best served by creating a typology of data variants based off of usage, how the material is treated, and ultimately how many levels the data is “away from” its origin material.

---

[184] Ibid.

[185] Ibid., 28.

[186] Matthew Stanley, “Where Is That Moon, Anyway? The Problem of Interpreting Historical Solar Eclipse Observations” Gitelman, *“Raw Data” Is an Oxymoron*, 77.

[187] David Ribes and Steven J. Jackson, “Data Bite Man: The Work of Sustaining a Long-Term Study” ibid., 161.

[188] David Ribes and Steven J. Jackson, “Data Bite Man: The Work of Sustaining a Long-Term Study” ibid., 149.

Furthermore, if data can retroactively influence and shape its contexts, could it be argued that the same can be said for their relationship with the digital archive? Are metadata standards influencing how we approach and conceptualise data? Is data dictating how we read and interact with it? Problems emerge because what data are is being delimited by its surrounding architecture; with the term data used and re-used in the context of the dataset, the database, and in relation to data of various levels of processing. How do we maintain an awareness of which data is what data? And how to accommodate fluidity, a capacity for change, and account for the incorporation of material not identified as data at the time of its entry into the system to subsequently *become* data?

These differential interpretations of data can be conceived of through the vista of a myriad of philosophical, anthropological or sociological theories. Fundamentally, however, we appear to have a problem of language, with the term data having become grossly overdetermined and being no longer capable of clearly signifying what it is we or others talk about when we talk about data. Indeed it is likely a combination of these phenomena and so, much like the term itself, our explanations of the reasons have the potential to become myriad. Irrespective of how one elects to theorise why data is or has become so problematic, there is extensive confusion discernible between the gradations of data one encounters in a digital environment, and this needs to be redressed. We have data at the level of input, we have data that has been cleaned, we have data that is present but not encoded, we have data that only becomes data when it is excised from its original context and situated in another context, but for all of these different phases or evolutions of data, we have only one term; and thus we have minimal opportunities (aside from context within an article, discipline specific intuition, explanatory footnotes etc.) to deduce what type of data is being referred to, or what has been done to "this data" to make it different to "that data." Hong Liu refers to two data revolutions, the first data revolution being "the fusion of data and scientific research" and the second being "the *emergence of big data*."[189] With this in mind, we will first turn our attention to data as it is and has been defined, then to the implications of the merger of data and scientific research (particularly for the humanities), and finally, to the concept of big data.

Borgman identifies four main approaches to the definition of data, these are:

1. Definition by Example
2. Operational Definitions
3. Categorical Definitions
4. Conceptual Distinctions[190]

It will be useful to expand and elaborate on these broad delineations for the purposes of our study. The following can be considered an overview of the various conceptions of data available to the researcher, and the communities in which these conceptions operate.

1. Data defined by archival principal or its treatment within the context of an archive.
2. Data as user defined/ Data defined by usage.

[189] Hong Liu, "Philosophical Reflections on Data," Procedia Computer Science, 1st International Conference on Data Science, ICDS 2014, 30 (January 1, 2014): 62, doi:10.1016/j.procs.2014.05.381.

[190] Borgman, "Big Data, Little Data, No Data," 19–26.

3. Data defined by origin.
4. Defined by treatment/ how it was captured.
5. Data as ambiguously defined, data as a synonym for input.
6. Data and the humanities.
7. Big data.

2.2.1. Data defined by archival principal or its treatment within the context of an archive.

This falls under Borgman's category of "Operational Definitions" with Borgman observing that "The most concrete definitions of data are found in operational contexts." [191] Operational definitions of data are pragmatic and, for the most part, discipline specific, which concordantly means that the problems encountered on a discipline-specific small scale environment will be magnified on an interdisciplinary level. These definitions of data are not so much definitions *per se*, but archival principles. Concordantly, they are often unclear in relation to what is and is not data; and arguably they also delimit the re-interpretability of the data by presenting it in the context of a specific database or dataset. Further still because it is the entity's placement within a database that renders it data, there is a high degree of contextual input, with the metadata taking on a prominent role in the assignation of data *as data*.

One of the most widely known and used principles for data archiving that operates on archival principles is the *Reference Model for an Open Archival Information System* (OAIS) (Consultative Committee for Space Data Systems 2012). Borgman observes that while these recommendations originated in the space sciences community, they have also been adopted in sciences and social sciences as a guideline.[192] The relevant section from the OAIS:

> *Data*: A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing. Examples of data include a sequence of bits, a table of numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen.[193]

The OASI distinguishes data from information. That said, however, the above OAIS definition arguably conflicts with Machlup and Mansfield's DIKW model, which we already known to be problematic and oversimplified because, as Borgman notes, the "tripartite division of data, information, and knowledge [...] oversimplifies the relationships among these complex constructs."[194] Of interest too in this respect is Borgman's later observation that the OAIS uses data not as a noun, but as a modifier: "The OAIS Reference Model uses *data* as a modifier: dataset, data unit, data format, database, data object, data entity, and so on, while defining data in general terms with examples."[195] The question as to whether this grammatical shift represents an improvement is difficult to answer; on the one hand it

[191] Ibid., 20.

[192] Ibid., 20.

[193] OAIS Reference Model, qtd. in ibid., emphasis in original.

[194] Ibid., 17.

[195] Ibid., 20.

acknowledges the functional multiplicity of “data,” but on the other hand by disassociating the term “data” from the nominal the term is further removed from engagement and disassociated from the active. As a modifier, the implication is that “data” is somehow less active or is in some way subordinate to the head term or noun, whatever that may be.

In many cases the data is curated to reflect the objectives of its host institution, and so we can also forward the argument that data are entities that are sometimes defined by institution: “Museums, libraries, and archives navigate those [arbitrary] boundaries in determining what to collect. The roles of these ‘memory institutions’ overlap substantially. Each institution will represent and arrange its objects according to its mission.”[196]

Models such as DIKW create misleading and misrepresentative impressions about the supposed distinctions between the facets of DIKW, leading to situations where data is defined *by definition* (“Because we have DIKW, data must equal ‘x-value’”), or for example, in the passage by Hong Lui that follows, data is defined by rote:

> Data becomes the necessary basis of the information, knowledge and wisdom, and penetrates into them, forming to a system of “data, information, knowledge and wisdom (DIKW)”. The number in the data is very abstract, and only when adding background, and transforming the number into a quantity with a specific meaning, the data is capable of turning into the valuable information. Therefore, from data to information it is necessary to clean up the data, which is the central task of data mining. Only from information can rules and laws be refined to generate knowledge, while the refining process is composed of the induction, deduction, and so on. Under the framework of DIKW system, according to the process of scientific research, from the acquisition of raw data, to the production of derived data, and until the formation of knowledge data, the general course to form data and their roles can be addressed.[197]

This passage is confusing but largely representative of the problems associated with conflating material and methodological approaches from the Sciences, Social Sciences and Humanities under the ungradiated umbrella-term “data.” The term “data” is employed too frequently, with consistently unclear and changeable referents. This creates semantic confusion around what data are, and moreover when specified data are referred to as a numerical (and abstract) entity, when this is not always the case, particularly in the humanities. Data are at times intimated as being already divorced from its native context, its pre-cleaning and post-cleaning state referred to using the same term (“data”) and so that which has been removed (the “dirt”) together with the approaches adopted by the cleaner, are left unacknowledged. Data is presented as synecdoche, with the partial data and larger data of which the partial is but a subset still nevertheless referred to simply as “data.” Therefore we are dealing with an entity whose sum total, part, and excised part are all referred to using the same term: data.

[196] Ibid., 166.

[197] Hong Liu, “Philosophical Reflections on Data,” *Procedia Computer Science*, 1st International Conference on Data Science, ICDS 2014, 30 (January 1, 2014): 64, doi:10.1016/j.procs.2014.05.381.

2.2.2. Data as user defined/ data defined by usage.

The DDI (Data Documentation Initiative) is "a set of metadata standards for managing data throughout their life cycle."[198] The DDI was developed by the Inter-University Consortium for Political and Social Research (ICPSR) (amongst others) and it "allows contributors to determine what they consider to be their data."[199] This so-called definition is effected by inputting the proto-data into the database in a manner that accords with the DDI metadata's specifications. DDI present data as something that is user defined then, a treatment that accords with the data as an entity of speculative value, and with Raley's idea of data as performative; anything can *be data* once it is entered into the system *as data*. Borgman:

> The DDI is widely used in the social sciences and elsewhere for data description but does not define data per se. The DDI metadata specifications, which are expressed in XML, can be applied to whatever digital objects the DDI user considered to be data.[200]

From this we can conclude that in addition to data being performative, metadata is also performative, having the potential to situate proto-data as data proper.

Data is also defined in terms of that which is considered convenient or practical. Borgman refers to this categorical approach as "grouping them [data] in useful ways," and it is discernible in "operational and general research contexts"[201]: "Data archives may group data by degree of processing, for example. Science policy analysis may group data by their origins, value, or other factors."[202] This approach is a variant on the preceding category that saw data as user defined. In this instance, data are defined on the basis that which is considered practical or pragmatic at the time, deriving its status as data because it serves a purpose that aligns with the particular usage requirements of the project, collection, or archive. Thus, grouping can itself be considered as a process or an act of curation; one that is, or has the potential to be, idiosyncratic and highly flawed.

A related approach to the act of defining data by user or usage comes in the form or approaches that see data defined by topic/ category/ profession. Eg. Healthcare/ biomedicine. With this comes a blurring of the distinctions between professions and categories that are particularly problematic when it comes to consent issues. This includes de-identified personal data. Defining data by topic, category or profession is also problematic when it comes to facets of the humanities (and other research disciplines, the medical humanities, for example) that overlap.

2.2.3. Data defined by origin.

[198] Borgman, "Big Data, Little Data, No Data," 20, emphasis in original.

[199] Ibid., 21, emphasis in original.

[200] Ibid., 20, emphasis in original.

[201] Ibid., 21.

[202] Ibid.

Hong Lui observes “the instrumental characteristics of data was always the epistemological basis of the data utilization.”[203] This reflects the categories promoted by the US National Science board (NSB) which advocates three core categories of data that focus on classifying data on the basis of how it was recorded, in other words, by origin. These categories are Observational data, Computational data, and Experimental data, with a fourth sub-category (or meta-category) in the form of “derivative data” brought about by “processing and curatorial activities.”[204] Lui develops on these three categories noting that

> Data usually has the following types, observational data, experimental data, theoretical data, simulated data, statistical data, and big data. [...] With the advancement of modern science and technology, the form and connotation of data are also changing and developing, while in addition to the observational data, the experimental data, theoretical data, simulated data, statistical data, and digitized graph, table, text, etc. are also important ingredients of data.[205]

Lui’s argument here is an example of how easily grossly differential typologies of data can be collated and placed in situ. Big data is not the same kind of data as observational data, and so on. To place them in situ because of a common noun (“data”) creates an impression of comparability that is misrepresentative. In addition, Lui conflates elements that contribute to smaller scale conceptions of data: “graph, table, [and] text” are not necessarily “ingredients of data,” but can in fact be data proper. And finally, while the NSB, as Borgman notes, “is intended to reflect data used in the sciences, social sciences, and technology. [and] the humanities, arts, medicine, and health are outside the remit of NSB, these data categories are [also] intended for use in those domains."[206] A question remains as to whether or not these categories are useful (or even relevant) in the context of the humanities, and whether they may need to be adapted to suit certain of the idiosyncrasies of humanities research. Further still, this process of NSB classification by origin does not actively flag the refinement, processing, or scrubbing the data has undergone to transition from its native proto-data state to the state of having been captured and undergone datafication.

Borgman distinguishes between Sciences and Social Sciences & Humanities throughout her analyses of data. In the context of this study this distinction is a useful reminder that the practices surrounding data collection and preparation in the Sciences etc. are different (and necessarily different) to those in the Humanities. That said, however, why does US National Science Board intend their data categories to be used in domains outside of their remit?[207] Furthermore, there is an (unproven) assumption that interdisciplinary questions will crops data silos. In addition, the idea that data can be disciplinary is somehow tied to set disciplines is a fabrication. Having humanities researchers too readily adopted practices and protocols from the Sciences without thinking about the repercussions of these methodological approaches is problematic, and will be addressed later in this WP. For now, however, it is enough to note that this crossover between computer science practice,

[203] Liu, “Philosophical Reflections on Data,” January 1, 2014, 61.

[204] “NSB-05-40, Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century,” n.d.

[205] Liu, “Philosophical Reflections on Data,” January 1, 2014, 61–62.

[206] Borgman, “Big Data, Little Data, No Data,” 23.

[207] Ibid.

science, and the humanities appears to be one directional, with computational and scientific methods spilling into the humanities without an equal but opposite response that sees humanities methodologies start to spread into/ influence computer science/ the sciences in return. This is a factor noted by Edmond in *Will Historians Ever Have Big Data.*

2.2.4. Data defined by treatment/ how it was captured.

This sees data defined by what researchers do to the data to make it data. The major example of this comes in the form of NASA's Earth Observing System Data Information System (EOS DIS)[208] wherein, as Borgman notes, "Data with common origin are distinguished by how they are treated."[209] The EOS DIS is reproduced below:

Data Level	**Description**
Level 0	Reconstructed, unprocessed instrument and payload data at full resolution, with any and all communications artifacts (e.g., synchronization frames, communications headers, duplicate data) removed. (In most cases, the EOS Data and Operations System (EDOS) provides these data to the data centers as production data sets for processing by the Science Data Processing Segment (SDPS) or by a SIPS to produce higher-level products.)
Level 1A	Reconstructed, unprocessed instrument data at full resolution, time-referenced, and annotated with ancillary information, including radiometric and geometric calibration coefficients and georeferencing parameters (e.g., platform ephemeris) computed and appended but not applied to Level 0 data.
Level 1B	Level 1A data that have been processed to sensor units (not all instruments have Level 1B source data).
Level 2	Derived geophysical variables at the same resolution and location as Level 1 source data.
Level 3	Variables mapped on uniform space-time grid scales, usually with some completeness and consistency.
Level 4	Model output or results from analyses of lower-level data (e.g., variables derived from multiple measurements).

[210]

The EOS DIS is perhaps one of the most functional definitions of data available because it not only acknowledged the levels of processing material undergoes to become data, but tiers this scrubbing or cleaning process, therein acknowledging that some material undergoes more extensive modification than others, and maintaining traceability to the source context or environ wherein the “native data” was extracted from.

[208] “NASA EOS DIS Data Processing Levels,” n.d., https://science.nasa.gov/earth-science/earth-science-data/data-processing-levels-for-eosdis-data-products.

[209] Borgman, “Big Data, Little Data, No Data,” 21.

[210] “NASA EOS DIS Data Processing Levels.”

That said, the above table is not without its problems. Firstly, the table is incomplete because of the presence of a level that precedes “level 0”; data that precedes “level 0” is referred to in passing by Borgman as “native data,” a phrase that I find particularly acute and useful because it retains emphasis on the importance of context apropos data, whether that context be “native” or in the form of the context(s) it acquires when it transitions from a native to a non-native environment: "Some scientists want level 0 data or possibly even more native data without the communication artefacts removes so they can do their own data cleaning."[211] Second, while the distinctions between levels are relatively explicit here, they only pertain to the onset of the research, the point where data is *gathered* at the onset of the study; thereafter the data is considered raw, until it is subjected to further processing:

> Although NASA makes explicit distinctions between raw and processed data for operational purposes, *raw* is a relative term, as others have noted [...] What is 'raw' depends on where the inquiry begins. To scientists combining level 4 data products from multiple NASA missions, those may be the raw data with which they start. At the other extreme is tracing the origins of data backward from the state when an instrument first detected a signal.[212]

Further still, despite the usefulness of the term “native data,” even native data has been processed to a degree, and so one can also argue, as Borgman does, that "Identifying the most raw form of data may be an infinite regress to epistemological choices about what knowledge might be sought."[213]

It is important to note and take into account the distinction between quantitative data in the form of the NASA EOS DIS and qualitative data such as that outlined in YS Lincoln and EG Guba’s *Naturalistic Inquiry.*[214]

It remains to be seen as to how the NASA EOS DIS table would look had it been compiled for the purpose of humanities research. Interestingly, not one of the categories employed in the NASA EOS DIS has an analogous one in the humanities (aside from the rather loose concept of primary, secondary and tertiary sources), though that is not to say that a clear, lucid gradation of data that distinguishes how the material has been treated, or at least flags the fact that the data has been subjected to transformations, would not be beneficial for humanities researchers; a data “passport” that flags the processing, transformations and contextualisation the material has been exposed to, or expunged from. And again, the fact that the NASA EOS DIS distinctions only apply up to the point where analysis begins (at which point the data, irrespective of processing level, becomes “raw data” within the context of that inquiry) is itself problematic.

Whereas the NASA EOS DIS classifies data by level of processing and the NSB classifies by origin (of input), they are nonetheless comparable because, to varying degrees both

[211] Borgman, “Big Data, Little Data, No Data,” 22.

[212] Ibid., 26.

[213] Ibid., 27.

[214] Newbury Park, CA: Sage Publications, 1985.

represent datafication as the imposition of what Borgman classifies as "levels of data"[215] that reflect the degree of refinement or scrubbing the data has undergone. Even when it comes to a layout as seemingly precise as the EOS DIS, however, we still have to acknowledge the arbitrariness of what we're dealing with. The categories promoted by the NSB, or by NASA's EOS DIS can be compared to the SHOAH VHA's keyword database and a similar problem, articulated in the following passage by Borgman, can be identified across all three:

> No matter how sharp these distinctions between categories may appear, all are arbitrary to some degree. Every category, and name of category, is the result of decisions about criteria and naming. Even the most concrete metrics, such as temperature, height, and geo-spatial location, are human inventions. Similarly, the measurement systems of feet and inches, meters and grams, and centigrade and Fahrenheit reflect centuries of negotiation.[216]

While this argument is ontological and has the potential to take this research on a philosophical tangent that is anathematic to the pragmatic aims of the study, it is nevertheless worth keeping in mind that categories or taxonomies that appear neutral or objective *now*, may not appear so in the future; acknowledging that our research is the project of, and reflective of, specific cultural periods and research practices is important; particularly when it comes to capturing and reflecting the ambiguities of humanities research data. William Uricchio is particularly enlightening on this topic, and on the necessity for continual evaluation of the intersection between data, information architecture and humanities research. Uricchio observes:

> Data, the structure of the data set, models, software systems and interfaces all play determining roles in cultural production and, as such, are not only appropriate but increasingly important sites for humanistic inquiry. Their analysis requires not only new literacies but evaluative paradigms that in many cases have yet to be created.[217]

2.2.5. Data as ambiguously defined, data as a synonym for input.

Having data as an ambiguously defined entity can be deliberate, a purposeful decision not to "impose precise definitions" so that "data remains an ambiguous concept"[218] and, correspondingly, the organisations responsible for these ambiguous (non-)definitions can, as Borgman notes, "adapt to new forms of data as they appear."[219] This seems to be something of a back-footed approach that skirts the central issue of the need for a reconceptualization of what we talk about when we talk about data in favour of maintaining status quo while arguing that doing so facilitates innovation or new forms. Rather this is an approach that not perpetuates an ongoing problem, and in her above analysis, Borgman is perhaps trying to

[215] Borgman, "Big Data, Little Data, No Data," 24.

[216] Ibid., 26.

[217] William Uricchio, "Data, Culture and the Ambivalence of Algorithms," Mirko Tobias Schäfer and Karin van Es, *The Datafied Society. Studying Culture through Data*, 128.

[218] Borgman, "Big Data, Little Data, No Data," 21.

[219] Ibid.

put a positive spin on things by arguing that such an approach is a) deliberate and b) mindful of and responsive to innovation.

This also allows for and facilitates user-defined approaches to data, but it also arguably delimits users ability to identify data, because without a functional definition, proto-data may not be identified as data proper:

> Institutions responsible for managing large data collections should be explicit about what entities they handle and how, but few of these definitions draw clear boundaries between what are and are not data.[220]

Conceiving of data as ambiguously defined input is particularly popular among humanities researchers, with the classificatory approach discernible in the NSA and the detailed tiered approach of the NASA EOS DIS all but absent. Borgman makes the nice observation that "Distinctions between primary and secondary sources [in the humanities] are the closest analog [sic.] to raw and processed groupings in the sciences and social science."[221] To this we could add of course the tertiary sources or structural metadata provided by bibliographies, classification systems and indexes, and so on. But a more detailed tiered typology is not only viable, but potentially a very useful avenue to explore in relation to maintaining clarity regarding the modifications that have been enacted on the data used by researchers.

2.2.6. Data and the humanities.

Eef Masson makes the useful observation that humanities scholars largely "do not seek to establish unassailable, objective truths" and "instead […] approach their objects of study from interpretive and critical perspectives, acting in the assumption that in doing so they necessarily also preconstitute them."[222] This approach has consequences when it comes to crossover between humanities and the sciences, and crossovers with computer science in particular:

> with the introduction of digital research tools, and tools for data research specifically, humanistic scholarship seems to get increasingly indebted to positivist traditions. For one, this is because those tools, more often than not, are borrowed from disciplines centred on the analysis of empirical, usually quantitative data. Inevitably, then, they incorporate the epistemic traditions they derive from.[223]

Borgman identifies uncertainty as a key factor that distinguishes the way data is identified and treated in the humanities:

[220] Ibid., 20.

[221] Ibid., 27.

[222] Eef Masson, "Humanistic Data Research An Encounter between Epistemic Traditions," in Mirko Tobias Schäfer and Karin van Es, *The Datafied Society. Studying Culture through Data*, 25.

[223] Eef Masson, "Humanistic Data Research An Encounter between Epistemic Traditions," in ibid.

> An important distinction about data in the humanities is how uncertainty is treated in knowledge representation (Kouw, Van den Heuvel, and Scharnhorst 2013). Uncertainty takes many forms, whether epistemic, statistcal, methodological, or sociocultural (Peterson 2012). Ambiguity and heterogeneity are sources of uncertainty in historical records, for example. As scholars in the humanities apply technologies developed for other forms of inquiry, such as statistical tools and geographical information systems, they are caught in the quandary of adapting their methods to the tools versus adapting the tools to their methods. New tools lead to new representations and interpretations. Each field, collectively, and each scholar, individually, assesses how much uncertainty can be tolerates and what constitutes 'truth' in any inquiry. Implicit in research methods and in representation of data are choices of how to reduce uncertainty.[224]

Kouw, Van Den Heuvel, and Scharnhorst also acknowledge what they term the "highly ambiguous meaning of data in the humanities" and refer to the American Council of Learned Society's 2006 Commission on Cyberinfrastructure for the Humanities and Social Sciences titled *Our Cultural Commonwealth*.[225] This amalgamation of data-definition stands in contrast with the relatively clear tiered table provided in NASA's Earth Observing System Data Information System (EOS DIS) wherein "Data with common origin are distinguished by how they are treated."[226]

In her chapter on "Data Scholarship in the Humanities" Borgman observes that "Because almost anything can be used as evidence of human activity, it is extremely difficult to set boundaries on what are and are not potential sources of data for humanities scholarship. [...] Whether something becomes a source or resource for data in the humanities may depend on its form, genre, origin, or degree of transformation from its original state."[227] Whereas the structural architecture of NASA's EOS DIS makes explicit what processing entails, and also allows for the researcher to revert to a lower level of data, reverting to a pre-processing phase or to "native data," the metadata used in the case of humanities research is not as methodologically clear-cut, nor does it provide the opportunity for the researcher to access unfiltered "native data" in its original state or context, and thus proto-data that is ostensibly devoid of any perceived contextual interferences: "Metadata provide context, but the question of whose context is particularly contentious in the humanities."[228] Data for the humanities needs to be maintained in a native or near-native format, so as to facilitate the humanist's tendency to "[draw] on all imaginable sources of evidence."[229]

---

[224] Borgman, "Big Data, Little Data, No Data," 28.

[225] American Council of Learned Socieyties Commission on Cyberinfrastructure for the Humanities and Social Sciences, Our Cultural Commonwealth (2006), available at "ACLS American Council of Learned Societies | Home," *ACLS American Council of Learned Societies | Www.Acls.Org*, accessed May 22, 2017, http://www.acls.org.

[226] Borgman, "Big Data, Little Data, No Data," 21.

[227] Ibid., 166–67.

[228] Ibid., 171.

[229] Ibid., 161.

The issue of metadata's contextual and formative influence on data fuels what is referred to as the "*Surrogates versus Full Content*"[230] debate. The movement towards interoperability between disciplines in the humanities also contributes to the debate on the role, function, and influence of metadata on data proper:

> Formal mechanisms such as these [*Art and Architecture Thesaurus*, *Union List of Artist Names*, *Getty Thesaurus of Geographic Names*, *Cultural Objects Name Authority*] promote standardization and interoperability between systems. They provide the rigorous structure necessary for curatorial databases of museums, archives, and libraries.[231]

This presents data management and curation as a balancing act between structure, institutional rigor, and curatorial or documentation standards and the individuative nature of the materials that constitute (or have the potential to constitute) data within the humanities. Borgman notes "Humanities scholars rely heavily on unique material—letters, memos, treaties, photographs, and so on—that exist only in physical form."[232] Integrating/ incorporating/ accounting for and facilitating these unique artefacts entails incorporating uncertainty within rigorously structured systems; this dilemma over how to impart richness, ambiguity and interpretative uncertainty within a digital environment is a major impediment to our efforts at facilitating a big data research environment for humanities researchers.

This is further compounded by the fact that data conceived of as discipline-specific cannot be produced solely by employing discipline specific methods; a fact that is particularly problematic for humanities researchers as they integrate methodologies borrowed from computer science and the sciences. The disjunction between conceiving of data as discipline-specific when this same data fundamentally cannot be produced solely by employing discipline specific methods again displays something of a disconnect in the conceptualisation of the relationships that exist between data and origin, and between data and discipline: humanities data is not necessarily always produced solely by humanities research. On the contrary, the integration of methodologies borrowed from the sciences is nothing new, as evidenced by the history of quantification in literary science in the form of the structuring and formal representation of knowledge with vocabularies and ontologies, which has led to linked open data, semantic web, taxonomies & ontologies; and the precursors of big data in the historical sciences (census, agricultural statistics).[233]

Data is fundamentally inter-disciplinary, and even trans-disciplinary. Matthew Stanley, speaking in the context of astronomy, notes that "Astronomical data could not be produced solely by astronomical methods."[234] These additional necessary methods are idiosyncratic and difficult to anticipate or systematise; for example in the case outlines in Stanley's

[230] Ibid., 167.

[231] Ibid., 172.

[232] Ibid., 162.

[233] See: Konrad H. Jarausch, »Möglichkeiten und Probleme der Quantifizierung in der Geschichtswissenschaft«, in: Konrad H. Jarausch (Hg.), Quantifizierung in der Geschichtswissenschaft. Probleme und Möglichkeiten, Düsseldorf: Droste 1976, S. 11-30.

[234] Matthew Stanley, "Where Is That Moon, Anyway? The Problem of Interpreting Historical Solar Eclipse Observations" Gitelman, *"Raw Data" Is an Oxymoron*, 84.

chapter on astronomy "The text could only become data through a proper knowledge of Latin grammar."[235] Each additional facet necessary for the creation of the data subjects the material to modification (transformation) that is not necessarily reversible, and to processes that are not always recorded, particularly in the case of what Stanley refers to as "psychological filters"; that is to say, psychologically motivated curation effects:

> The goal of all of these struggles was obtaining a number: the secular acceleration, which would then modify the equations of the moon's motion. To get this number, one needed other numbers: the time and place of ancient eclipses. But this data only existed once it had been passed through textual, historical, and psychological filters. Each filter could be used positively or negatively, to either exclude a record from reliability or to detect reliable records.[236]

Subject specific data cannot be produced using subject specific methods then, and there are deficiencies in understanding the epistemological implications of turning to these methods. Furthermore, within disciplines, approaches to "cleaning"/ "scrubbing" or modifying data differ greatly, with not enough transparency surrounding the reasons for these procedures, or their long-term consequences (given that the cleaning, the decisions behind it and the methodologies adopted are frequently left out of the material produced).

Acknowledging the impossibility/ improbability of researchers having all the requisite skills necessary to take data from these materials is necessary, as is the need to be sensitive to the problems introduced into the project by incorporating non-discipline specific methods. The reason this latter task has not been widely or actively pursued is likely down to unfamiliarity with the epistemological implications: problems inherent to these methodologies, along with a worrying lack of research within these disciplines themselves regarding the epistemologies of their own methods and protocols. This ignorance towards the epistemologies of methods outside one's discipline is not surprising; if you are unfamiliar with the basic principles of another discipline, you are likely to similarly ignorant of the epistemological implications of these same principles; these field is itself often a-sub-facet/ discipline specific field of specialisation.

Kouw, Van Den Heuvel, and Scharnhorst look at various approaches to the classification of uncertainty and this overview, together with their analysis of the different approaches taken and taxonomies adopted could be usefully transferred to our attempts to do the same with data. They cite Brugnach et al (2008) and their idea of a

> relational concept of uncertainty, which 'involves three elements: 1. an object of perception or knowledge [...] 2. one of more knowing actors [...] for whom that knowledge is relevant; and 3. different knowledge relationships that can be established among the actors and the objects of knowledge.' [...] In this framework, there may be three causes of uncertainty. First, we may be dealing with systems whose behaviour can only be predicted to some extent. Second, we may have

[235] Matthew Stanley, "Where Is That Moon, Anyway? The Problem of Interpreting Historical Solar Eclipse Observations" ibid.

[236] Matthew Stanley, "Where Is That Moon, Anyway? The Problem of Interpreting Historical Solar Eclipse Observations" ibid.

> incomplete knowledge of the system in question. Third, there may be different or even incompatible frames of reference for the system in question.[237]

Their criticism of Brugnach et al's system raises an important point that may be relevant to the creation of a new conceptualisation of data, specifically the fact that they do not differentiate between epistemic and ontic uncertainty.[238] There is perhaps an opportunity here for us when we are flagging or categorising data to differentiate between epistemic and ontic data.

Kouw, Van Den Heuvel, and Scharnhorst advocate the benefits of "uncertainty as a source of knowledge production"[239] and indeed it is areas of uncertain, proto-, or speculative value that harbour heretofore unidentified data (which in itself recalls point 2.15: Data as ambiguously defined, data as a synonym for input) that have the potential to yield rich scholarly findings. Such an approach is not without difficulties, particularly when approaching it through a digitized environment, and one avenue the authors suggest is to

> [develop] a closed system of explanations that leaves a *particular* (and potentially *restricted*) arena for knowledge production. An emphasis on both [epistemic and uncertain] forms of knowledge is necessary, which implies a balancing act between relying on firm epistemic grounds and carefully broadening one's scope so that new avenues of knowledge can be explored."[240]

Bruno Latour "proposes to recalibrate, or realign, knowledge with uncertainty, and thereby remains open to a productive disruptive aspect of uncertainty."[241] This is potentially useful because, as Kouw, Van Den Heuvel, and Scharnhorst point out, "uncertainty is often explained as a lack of knowledge, or as an aspect of knowledge that implies a degree of unknowability. Such interpretations can result in commitments to acquire more information about a particular situation, system, or phenomenon, with the hope of avoiding further surprises."[242] This is an anathema to the promotion of a rich data environment for humanities researchers, and it is important to examine "how uncertainty can be a source of knowledge that can disrupt categories that provide epistemological bearing."[243] We need to facilitate a conceptualisation of data that allows for and incorporates this, even if it just through a tiered taxonomy or tagging system.

---

[237] PAUL WOUTERS et al., eds., *Virtual Knowledge: Experimenting in the Humanities and the Social Sciences* (MIT Press, 2013), 95, http://www.jstor.org/stable/j.ctt5vjrxn.

[238] Ibid.

[239] Ibid., 90.

[240] Ibid.

[241] Matthijs Kouw, Charles Van Den Heuvel, Andrea Scharnhorst, "Exploring Uncertainty in Knowledge Representations: Classifications, Simulatons, and Models of the World" in PAUL WOUTERS et al., eds., Virtual Knowledge: Experimenting in the Humanities and the Social Sciences (MIT Press, 2013), 89, http://www.jstor.org/stable/j.ctt5vjrxn.

[242] Matthijs Kouw, Charles Van Den Heuvel, Andrea Scharnhorst, "Exploring Uncertainty in Knowledge Representations: Classifications, Simulatons, and Models of the World" in ibid.

[243] WOUTERS et al., Virtual Knowledge, 2013, 90.

As Presner observes, the SHOAH VHA's approach to uncertainty within the digital archive was to tag the material as "'indeterminate data' such as 'non-indexable content.'"[244] Presner's account of this is worth quoting in full, as it not only raises a number of concerning ethical issues, but it highlights the real difficulties facing researchers when it comes to encoding particularly difficult, charges or rich materials into a dataset:

> 'Indeterminate data' such as 'non-indexable content,' must be given either a null value or not represented at all. How would emotion, for example, need to be represented to allow database queries? While certain feelings, such as helplessness, fear, abandonment, and attitudes, are tagged in the database, it would be challenging to mark-up emotion into a set of tables and parse it according to inheritance structures (sadness, happiness, fear, and so forth, all of which are different kinds of emotions), associative relationships (such as happiness linked to liberation, or tears to sadness and loss), and quantifiable degrees of intensity and expressiveness: weeping gently (1), crying (2), sobbing (3), bawling (4), inconsolable (5°. While we can quickly unpack the absurdity (not to mention the insensitivity) of such a pursuit, there are precedents for quantified approaches to cataloguing trauma [...] Needless to say, databases can only accommodate unambiguous enumeration, clear attributes, and definitive data values; *everything else is not in the database.* The point here is not to build a bigger, better, more totalizing database but that database as a genre always reaches its limits precisely at the limits of the data collected (or extracted, or indexed, or variously marked up) and the relationship that govern these data. We need narrative to interpret, understand, and make sense of data.[245]

Presner gestures towards rethinking the database as a genre: "I wonder how we might rethink the very genre of the database as a representational form"[246] specifically regarding representations of "the indeterminate […] Such a notion of the archive specifically disavows the finality of interpretation, relished in ambiguity, and constantly situates and resituates knowledge through varying perspectives, indeterminacy, and differential ontologies."[247] How could we rethink the database as a genre in this manner? How to we incorporate disavowal, ambiguity, perspectivism?

Petersen's uncertainty matrix (see below), perhaps in combination with an adapted version of the NASA EOS DIS, is as a potential model for a tiered typology of data for the humanities that would allow for the recognition (again, on a tiered level) of material whose content is uncertain or multivalent:

[244] Presner, in Fogu, Claudio, Kansteiner, Wulf, and Presner, Todd, *Probing the Ethics of Holocaust Culture*, History Unlimited (Cambridge: Harvard University Press, 2015), http://www.hup.harvard.edu/catalog.php?isbn=9780674970519.

[245] Presner, in ibid.

[246] Presner, in ibid.

[247] Presner, in ibid.

UNCERTAINTY MATRIX			Level of uncertainty (from determinism, through probability and possibility, to ignorance)			Nature of uncertainty		Qualification of knowledge base (backing)			Value-ladenness of choices		
Location ↓			Statistical uncertainty (range+ chance)	Scenario uncertainty (range as 'what-if' option)	Recognized ignorance	Knowledge-related uncertainty	Variability-related uncertainty	Weak –	Fair 0	Strong +	Small –	Medium 0	Large +
Context		Ecological, technological, economic, social and political representation											
Expert judgement		Narratives; storylines; advices											
Model	Model structure	Relations											
	Technical model	Software & hardware implementation											
	Model parameters												
	Model inputs	Input data; driving forces; input scenarios											
Data (in general sense)		Measurements; monitoring data; survey data											
Outputs		Indicators; statements											

248

2.2.7. Big data

Big data is not a new concept, as van Es and Schäfer remind us: "Optimistic reporting about 'big data' has made it easy to forget that data driven practices have been part of the emerging information society since the nineteenth century (Beniger 1989; Porter 1996; Campbell-Kelly 2003)."[249] Moreover, and this is important, van Es and Schäfer attribute the problems currently being addressed in relation to our increasingly "datafied society" as having their origins not in the emergence of "big data" but rather as a result of phenomena pertaining to

> "the computational turn" (Berry 2012; Braidotti 2013). [...] This turn began in the 1950s with the introduction of electronic computers and continues unabated today. It concerns the datafication of everything: all aspects of life are now transformed into quantifiable data (Mayer-Schönberger & Cukier 2013).[250]

Rosenthal identifies this and is nicely articulate on the particular problem this poses to the evolution of literary criticism, arguing that we can read "the relationship between conventional literary criticism (close reading, microanalysis) and data-driven literary criticism

[248] Arthur C. Petersen, Simulating Nature: A Philosophical Study of Computer-Simulation Uncertainties and their Role in Climate Science and Policy Advice, Uitgeverji Maklu (2006), reprinted in WOUTERS et al., *Virtual Knowledge*, 2013, 97.

[249] Karin van Es and Mirko Tobias Schäfer, "Introduction" in Mirko Tobias Schäfer and Karin van Es, *The Datafied Society. Studying Culture through Data*, 13.

[250] Ibid.

(distant reading, macroanalysis) in terms of a relationship between narrative and data that has been going on for over a century."[251]

It is worth remembering and reiterating first of all that the concept of "big data" is not so new: census data exist for more than 2000 years, socio-demographic and economic big data have been collected for more than 200 years. Second, many of the problems being encountered and addressed are not so much the result of the rise of "big data" but rather ongoing fallout pertaining to the computational turn in contemporary society. Accordingly,

> "Digital humanities" is merely the nom de guerre of the computational turn in the humanities. Datafication and computerization will come to affect all research agendas and inform the skill sets of students and scholars alike. [252]

Lipworth is apprehensive as to whether big data can live up to the expectations others have of it and whether it can return on the investments made in it; her reluctance to "buy into" big data are the result of unresolved ethical and epistemological issues: "the epistemological issues raised by big data research have important ethical implications and raise questions about the very possibility of big data research achieving its goals."[253] Further still, "the scientific literature on big data research is replete with articles describing unresolved technical and scientific barriers to big data research"[254]

Lipworth cautions against investing wholeheartedly in big data research, observing that the level of commitment to big data is not reflective of a fully functional, unproblematic system:

> The degree of scientific, medical, political, and commercial commitment to big data biomedical research might give the impression that its processes and systems were well-established and morally, politically, and economically clear-cut, sustainable, and non-problematic.[255]

The misapprehensions made in relation to data are also made in relation to big data, we can return to Gruber Garvey's assertion, noted earlier, that "[big] data will out."[256]

> First, Big Data is presumed to have the inherent "authority" to speak for itself [...] Kate Crawford concurs: "Data and data sets are not objective; they are creations of human design. We give numbers their voice, draw inferences from them, and define their meaning through our interpretations. Hidden biases in both the collection and analysis stages present considerable risks" (2013: para. 2). We can add that algorithmic tools are, by nature, opaque to most researchers.[257]

[251] Rosenthal, "Introduction," 4.

[252] Mirko Tobias Schäfer and Karin van Es, *The Datafied Society. Studying Culture through Data*, 15.

[253] Lipworth et al., "Ethics and Epistemology in Big Data Research," 1, abstract.

[254] Ibid., 5.

[255] Ibid., 3.

[256] Ellen Gruber Garvey, "'facts and FACTS:' Abolitionists' Database Innovations," Gitelman, *"Raw Data" Is an Oxymoron*, 90.

[257] Karin van Es, Nicolàs López Coombs & Thomas Boeschoten, "Towards a Reflexive Digital Data Analysis" Mirko Tobias Schäfer and Karin van Es, *The Datafied Society. Studying Culture through Data*, 172.

Nick Couldry refers to big data as a "myth"[258] and Lui observes that

> big data breaks through the research boundary between the natural science and social science, establishing the commensurability of data, and bridging the resource sharing of different disciplines via data.[259]

This bridging is, of course, replete with epistemological and ethical challenges, which is why it's particularly important to be very clear on how we're approaching and dealing with data.

Big data for the humanities must try to acknowledge and take ongoing measures to alert its users to oligoptic:

> The current wealth of data can tempt the humanities researcher to adopt new forms of empiricism. However, data are not natural phenomena, but always exist within a particular social context. It is all too easy to lose sight of the fact that "all data provide oligoptic views of the world: views from certain vantage points, using particular tools, rather than an all-seeing, infallible God's eye view" (Kitchin 2014b: 4).[260]

Wendy Lipworth notes the lack of consensus regarding a definition for "big data":

> While there is no agreed definition of "big data," the phrase is generally used to refer to a "new generation of technologies and architectures, designed to extract value from large volumes of a wide variety of data by enabling high-velocity capture, discovery and analysis" (Costa 2014, 434).[261]

Karin van Es and Mirko Tobias Schäfer are more straightforward in their description not of the definition, but of the characteristics of big data: "In lieu of an illustrative metaphor, the label 'big data' is used to describe a set of practices involving the collection, processing and analysis of large data sets."[262] As is well known, these large data sets have the following five features: "Big Data has five characteristics: volume, velocity, variety, veracity and value."[263] The characteristics of Big Data are accumulating so rapidly that even recent publications may be out of date regarding its characteristics. For example, in a publication from 2014 Liu elaborates on four of these characteristics, because the fifth one (veracity) had not been established yet:

[258] Nick Couldry, "The Myth of Big Data" ibid., 237.

[259] Liu, "Philosophical Reflections on Data," January 1, 2014, 63.

[260] Karin van Es, Nicolàs López Coombs & Thomas Boeschoten, "Towards a Reflexive Digital Data Analysis" Mirko Tobias Schäfer and Karin van Es, *The Datafied Society. Studying Culture through Data*, 171.

[261] Wendy Lipworth et al., "Ethics and Epistemology in Big Data Research," *Journal of Bioethical Inquiry*, March 20, 2017, 1., doi:10.1007/s11673-017-9771-3.

[262] Mirko Tobias Schäfer and Karin van Es, *The Datafied Society. Studying Culture through Data*, 13.

[263] See International Conference on Big Data and Internet of Thing (BDIOT2017), http://www.bdiot.org/cfp.html

> (i) volume (the huge amount of data); (ii) variety (the diverse and complex characters); (iii) velocity (the generating speed is fast and rapid); and (iv) value (the economic and scientific values); [264]

Again however we are faces with a theoretical conundrum: how can "data" have "no truth. Even today, when we speak about data, we make no assumptions about veracity"[265] when one of the characteristics of "big data" is "veracity"? How can data be "pre-analytical, pre-factual"[266] yet "big data" be at the same time both a "myth"[267] and something that can supposedly "speak for itself"?[268]

2.2.8. Summary: Data as synecdoche; data as synonym for input.

2.2.8.i. Same name, different data.

Data are teleologically flexible, epistemologically variable, arbitrary and categorically indistinct; data sources are also arbitrary and indistinct, particularly in the humanities. Data also contain errors and unintelligible material. Data are of speculative value, and if curated correctly, any facet of the native material input can be identified as data at any point. Furthermore, the integration of data and computer science in humanities research represents a crossover of methodological and ideological approaches. Humanities scholars' "methods tend towards idiographic explanation [...] As they work with larger amounts of data, nomothetic explanation may become more feasible, enabling the same questions to be explored across more contexts."[269]

There is certainly an awareness that loose and overlapping definitions of data can bring about problems, as Borgman makes this explicit in *Big Data, Little Data, No Data*:

> Data are most often defined by example [...] Lists of examples are not truly definitions because they do not establish clear boundaries between what is and is not included in a concept.[270]

Despite this, there is little real sense of how to tackle this problem, aside from recognising that it exists. Indeed, literature on the problems that are concordant with data tend to be discipline specific, and aside from Borgman's *Big Data Little Data*, there are no overarching studies (that I can find) that seek to acknowledge the extent of this problem, or that work to address it and propose solutions.

---

[264] Liu, "Philosophical Reflections on Data," January 1, 2014, 62.

[265] Rosenberg, "Data before the Fact," in ibid., 37.

[266] Rosenberg, "Data before the Fact," in Gitelman, "Raw Data" is an Oxymoron, 18.

[267] Nick Couldry, "The Myth of Big Data" ibid., 237.

[268] Karin van Es, Nicolàs López Coombs & Thomas Boeschoten, "Towards a Reflexive Digital Data Analysis" Mirko Tobias Schäfer and Karin van Es, *The Datafied Society. Studying Culture through Data*, 172.

[269] Borgman, "Big Data, Little Data, No Data," 162.

[270] Borgman, "Big Data, Little Data, No Data," 19.

The same term (data) is used for material captured from vastly different environments and using different methods; in other words, different data, same name. The same term is also used to refer to material that has undergone different levels of processing.

> Data not only becomes the basis of almost all disciplines involved, affects the change of scientific research paradigm, but also closely connects with all life and the environment of human beings. Data has become a kind of an unavoidable and inseparable language code.[271]

Perhaps as Presner intimates, the term data promises a "data sublime,"[272] a Platonic ideal that cannot live up to the expectations of its users who approach data expecting a totality that is not there. As discussed in Section 1, a tiered approach to data that accommodates data processing levels, or that acknowledges data transformations or recontextualisation and differentiates between different types of data (for example, proto-data, pre-data, pre-processed data, native data, raw data, and thereafter data of various levels of processing) would make for a more transparent research environment.

In addition scholars have long acknowledged the issues surrounding data cleaning and processing, particularly in the Sciences, but there appears to be lack of material addressing, acknowledging and accounting for data processing in the humanities. What also needs to be addressed is whether the cleaning data undergoes is (or should be) reversible in the way material recorded using NASA's EOS DIS is, where a researcher can opt for data at levels between 0 and 4, or even further back than the 0 phase and opt instead for native data (level pre-0 data?). These machinations and re-shapings or recontextualisation of data are under-acknowledged, rarely explained or justified, and often not reversible.

2.2.8.ii. Code, codework, and algorithms.

Aside from problems defining data, however, we also have ethical and epistemological issues surrounding the use of algorithms, coding, and the incorporation of methodologies and technologies from the sciences into humanities research. There are misconceptions surrounding our understanding of data, code and codework, and of the relationships that exist (or could exist) between narrative, data and code. These misconceptions centre on presumptions of objectivity and of totality, and confusion arises because the term data is used with abandon. The following is an astute summary of how software technology is being approached/ overlooked when it comes to conceptualizing on data and digital archives.

> code and codework are all too often treated as an invisible hand, influencing humanities research in ways that are neither transparent nor accounted for. The software used in research is treated as a black box in the sense of information science—that is, it is expected to produce a certain output given a certain input—but at the same time it is often mistrusted precisely for this lack of transparency. It is also often perceived as a mathematical and thus value neutral and socially inert

[271] Liu, "Philosophical Reflections on Data," January 1, 2014, 60.

[272] Presner, in Fogu, Claudio, Kansteiner, Wulf, and Presner, Todd, *Probing the Ethics of Holocaust Culture*.

> instrument; moreover, these two seemingly contradictory perceptions need not be mutually exclusive.[273]

Has the integral role change plays in code development been acknowledged (enough/ at all) or incorporated enough into decisions made regarding fitting software to humanities projects. Is there perhaps a reluctance towards change in methodology of documentalist/ archival origin that is hindering the potentiality for more innovative interpretations/ developments? Potentially important would be the need to elaborate more on how coding in the humanities is distinct from other disciplines? The problem here however is that "it is not common practice throughout the humanities to publish code."[274]

he chapter by Joris J. van Zundert, Smiljana Antonijević, Tara L. Andrews, "Black Boxes and True Color—A Rhetoric of Scholarly Code," provides "an ethnography of codework"[275] and ties in with the work of Passman & Boersma (*The Datafied Society*); Rieder & Röhle (also in *The Datafied Society*) is enlightening in this respect, particularly in respect to the consequences of ignoring the "black boxes of software":

> This lack of knowledge about what is actually taking place in these black boxes of software and how they are made introduces serious problems of evaluation and trust in humanities research. If we cannot read code or see the workings of the software as it functions, we experience it only in terms of its interface and its output, neither of which seem subject to our control. Yet code is written by people, which in turn makes it a social construct that embeds and expresses social and ideological beliefs of which it is also—intentionally or not, directly or as a side effect—an agent (cf. McPherson 2012). Since code is a more or less withdrawn (Berry 2011) or even covert, but non-neutral, technology, its users may unwittingly import certain methodological and epistemological assumptions in humanities research.[276]

So van Zundert, Antonijević, and Andrews draw attention to code as a social construct but also to the role of the programmer and to programming as a "narrative [that] is also part of an encompassing larger epistemological narrative":

> When a programmer writes software, the result is not merely a digital object with a specific computational function. It is a program that can be executed by a computer, but as so-called source code it is also a narrative readable by humans, primarily but not exclusively programmers (Hiller 2015). In the case of coding work in humanities research, that narrative is also a part of an encompassing larger epistemological narrative, comprising the research design, theory, activities, interactions and outputs. In the digital humanities context, the code-part of this narrative arises from a

[273] Joris J. van Zundert, Smiljana Antonijević, Tara L. Andrews, "Black Boxes and True Color—A Rhetoric of Scholarly Code," draft chapter, (unsure how to footnote this), 1.

[274] Joris J. van Zundert, Smiljana Antonijević, Tara L. Andrews, "Black Boxes and True Color—A Rhetoric of Scholarly Code," draft chapter, (unsure how to footnote this), 5.

[275] Joris J. van Zundert, Smiljana Antonijević, Tara L. Andrews, "Black Boxes and True Color—A Rhetoric of Scholarly Code," draft chapter, (unsure how to footnote this), 2.

[276] Joris J. van Zundert, Smiljana Antonijević, Tara L. Andrews, "Black Boxes and True Color—A Rhetoric of Scholarly Code," draft chapter, (unsure how to footnote this), 2.

> combination of the programmer's technichal skills, theoretical background knowledge (concerning both the humanities topic and computational modeling), and interpretations of the conversations she has had with collaborators both academic and technical. It follows that from an epistemic point of view, the practice of the programmer is no different from the practice of the scholar when it comes to writing (cf. Van Zundert 2016). Both are creating theories about existing epistemic objects (e.g. text and material artifacts, or data) by developing new epistemic objects (e.g. journal articles and critical editions, or code) to formulate and support these theories.[277]

van Zundert, Antonijević, and Andrews refer to code as "narrative" and argue that, like all narrative, code is not objective; rather they forward "the idea of code as an interpretative narrative."[278] This ties in nicely with the work of Rieder and Röhle, who effectively articulate what is essentially the same argument:

> Because for any experienced programmer, code may well be the medium of expression but, just like a writer attempts to say something through language, the meaning expressed through programming is functionality; and while the two cannot be fully separated, programmers and computer scientists generally reason on a conceptual level that is certainly circumscribed by the requirements of mechanical computation – what one of us has called the 'shadow of computation' (Rieder 2012) – but expressible in various forms, from systematized vocabulary and conversation to flowcharts and, more often than not, mathematical notation.[279]

The influence of code and codework is often overlooked or undervalued because of a mistaken assumption that code is discreet from culture: "codework is shaped and influenced by its (social) context, which may positively or negatively influence the attitude and perception coders hold towards their work."[280]

> Blackboxing can thus be perceived as a process of enclosing the tumultuous complexity of epistemological and methodological dilemmas, controversies, compromises, and decisions visible in the *process* yet hidden in the *output* of knowledge production.[281]

[277] Joris J. van Zundert, Smiljana Antonijević, Tara L. Andrews, "Black Boxes and True Color—A Rhetoric of Scholarly Code," draft chapter, (unsure how to footnote this), 2.

[278] Joris J. van Zundert, Smiljana Antonijević, Tara L. Andrews, "Black Boxes and True Color—A Rhetoric of Scholarly Code," draft chapter, (unsure how to footnote this), 4.

[279] Bernhard Rieder, Theo Röhle, "Digital Methods: From Challenges to Bildung" Mirko Tobias Schäfer and Karin van Es, *The Datafied Society. Studying Culture through Data*, 115.

[280] Joris J. van Zundert, Smiljana Antonijević, Tara L. Andrews, "Black Boxes and True Color—A Rhetoric of Scholarly Code," draft chapter, (unsure how to footnote this), 10.

[281] Joris J. van Zundert, Smiljana Antonijević, Tara L. Andrews, "Black Boxes and True Color—A Rhetoric of Scholarly Code," draft chapter, (unsure how to footnote this), 6.

The need to: "examine scholarship in the making, and follow and reopen the dilemmas and controversies of the process of knowledge production before they get enclosed in the black box."[282]

Building on the work of Coyne (1995) and Berry (2014), van Zundert, Antonijević, and Andrews argue against the tendency to perceive code as unerring and objective:

> the internal structure and narrative of code ought not to be regarded as a mathematically infallible epistemological construct—although formal and mathematical logic is involved in its composition, just as logic has a natural place within rhetoric. If we consider code as an interpretative narrative rather than a mathematically discrete argument, it parallels humanities knowledge production in terms of theory and methodology. Code can thus inherit the multi-perspective problematizing nature and diverse styles of reasoning that are a particular mark of methodology in the humanities. From this perspective, different code bases represent different theories, each of which needs to show its distinctive, true colors in order to be adequately recognized and evaluated.[283]

van Zundert, Antonijević, and Andrews argue the case for a "poetics" of code, for code as *technē*, for code as something that is stylistic and author specific:

> This feel is part of a personal style of working and a personal 'poetics' of code, which is important to adhere to.[284]
>
> Every coder has a personal experience of *technē* that is essential to her methods.[285]
>
> Similar to writing, code and coding is also the interpretation and reinterpretation of theory; like any narrative or theory, code is not some neutral re-representation. Its author selects, shifts focus, expresses, and emphasizes.[286]

This conception of code, and in particular their adoption of terms familiar to humanities researchers (poetics, *technē*) makes codework analogically comparable to language-based narrative production and speaking about code in this way is an effective way to "humanise" it and stress the necessity for humanities researchers to develop familiarity with it that is fully attuned to its epistemological complexities. As Rieder and Röhle observe, "Only then can we assess the potentials, limitations and styles of reasoning held by the tools we integrate into

[282] Joris J. van Zundert, Smiljana Antonijević, Tara L. Andrews, "Black Boxes and True Color—A Rhetoric of Scholarly Code," draft chapter, (unsure how to footnote this), 6.

[283] Joris J. van Zundert, Smiljana Antonijević, Tara L. Andrews, "Black Boxes and True Color—A Rhetoric of Scholarly Code," draft chapter, (unsure how to footnote this), 4.

[284] Joris J. van Zundert, Smiljana Antonijević, Tara L. Andrews, "Black Boxes and True Color—A Rhetoric of Scholarly Code," draft chapter, (unsure how to footnote this), 13.

[285] Joris J. van Zundert, Smiljana Antonijević, Tara L. Andrews, "Black Boxes and True Color—A Rhetoric of Scholarly Code," draft chapter, (unsure how to footnote this), 13.

[286] Joris J. van Zundert, Smiljana Antonijević, Tara L. Andrews, "Black Boxes and True Color—A Rhetoric of Scholarly Code," draft chapter, (unsure how to footnote this), 14.

our research configurations."[287] At the same time, there are limits to how versed one can become in other disciplines, and it is important when approaching these issues to remain cognisant of Rieder and Röhle's reminder that, much like Borgman's observation regarding the potential for an "infinite regress to epistemological choices"[288] when it comes to identifying the "most raw" form of any given data, what it means to "understand" these methodologies is relative: "In all of these examples, we ask what 'understanding' a computational technique would mean."[289]

It is also necessary to talk about algorithms. Even within their own disciplines (eg. computer science) the epistemological issues surrounding algorithm use is unaddressed and under-acknowledged. In a way it seems as though what they are lacking is a humanist approach to thinking about their methodologies. Uricchio:

> the algorithm enabled by big data [...] stands between the calculating subject and the object calculated; it refracts the subject-centred world. Together algorithms and data filter what we have access to, produce our texts with unseen hands and unknown logics, and reshape our texts, rendering them contingent, mutable and 'personalized'.[290]

What's interesting about the above passage is that while it addresses the integration of data and algorithms, and the effects these have on filtering the material we have access to, it does not acknowledge the pre-cleaning, cleaning, and modifying of data to make it ready to be entered into the database as data; what is done to data (pre-data/ proto-data) to make data *data* is unaccounted for. Thus far I have yet to encounter a study that adopts a holistic and inclusive approach that acknowledges and incorporates these facets in their entirety. The majority (if not all) studies identify and address one problem area (eg. data definitions, or data cleaning and how it is not acknowledged or recorded, or the problems we have integrating data and narrative, or problems with code, or algorithms etc.) whereas really if we are to enact or initiate real change we need to be thinking about all of these factors, because they all connect.

Just as data is used as synonym (and synechdoche) for input, so too is algorithm. Uricchio's chapter in *The Datafied Society* is particularly strong on this:

> Algorithms are invoked as synecdoche when the term stands in for a sociotechnical assemblage that includes the algorithm, model, data set, application and so on. They reveal a commitment to procedure, formalizing social facts into measurable data and clarifying problems into models for solution. And they function as talismans when the term implies an 'objectifying' scientific claim. Indeed, one might step back and note that these three uses say much more about social anxieties and aspirations than they do about algorithms. How, for example, can one make a claim to 'objectivity' with an

[287] Bernhard Rieder, Theo Röhle, "Digital Methods: From Challenges to Bildung" Mirko Tobias Schäfer and Karin van Es, *The Datafied Society. Studying Culture through Data*, 115.

[288] Borgman, "Big Data, Little Data, No Data," 27.

[289] Bernhard Rieder, Theo Röhle, "Digital Methods: From Challenges to Bildung" Mirko Tobias Schäfer and Karin van Es, *The Datafied Society. Studying Culture through Data*, 115.

[290] William Uricchio, "Data, Culture and the Ambivalence of Algorithms," ibid., 134.

> authored protocol whose operations depend on the highly variable character and structure of a particular data set?[291]

Again, as before with the term "data," "algorithm" is a multi-valent term with different meanings in different disciplines, different contexts, and different research (and other) communities:

> A further twist appears in the form of disciplinary specificity. The valences of the term 'algorithm' differ in mathematics, computer science, governance, predictive analytics, law and in the culture at large, complicating crossdisciplinary discussion.[292]

Rieder and Röhle also note the influence algorithms have on interpretation, not just within the "black box" of a software program, but at a visual level in terms of how they influence network visualization and the output of visualization tools:

> A very similar argument can be made for the popular field of network visualization. It is again important to notice that the point and line form comes with its own epistemic commitments and implications, and graph analysis and visualization tools like Gephi (Bastian et al. 2009) further structure the research process. But where do we go from there? If we consider that graph theory still provides powerful and interesting means to analyse a data set, what would critical analytical practice look like? For example, how can we consider the layout algorithms that transform dimensional adjacency matrices4 into two-dimensional network diagrams? These artefacts interpose themselves as mediators because each algorithm reveals the graph differently, highlighting specific aspects of its structure, thus producing a specific interpretation.[293]

Uricchio is right then to argue that the algorithm "is now being deployed in ways that redefine long-held subject-object relationships and, in so doing, it poses some rather fundamental epistemological questions."[294]

2.2.8.iii. Ethics & Ethical Implications of Epistemological Misconceptions.

Ongoing concerns regarding the ethics and ethical implications of the computational turn are prevalent, to the point where van Schie, Westra, and Schäfer note the emergence of a new discipline specificallty concerned with this very problem: "Emerging new branches of humanities research dealing with the use of digital methods are raising questions about methods and ethics."[295] There is a focus on the discriminatory effect of digital methodologies:

[291] William Uricchio, "Data, Culture and the Ambivalence of Algorithms," ibid., 127.

[292] William Uricchio, "Data, Culture and the Ambivalence of Algorithms," ibid.

[293] Bernhard Rieder, Theo Röhle, "Digital Methods: From Challenges to Bildung" ibid., 117.

[294] William Uricchio, "Data, Culture and the Ambivalence of Algorithms," ibid., 125.

[295] Gerwin van Schie, Irene Westra & Mirko Tobias Schäfer, "Get Your Hands Dirty: Emerging Data Practices as Challenge for Research Integrity" ibid., 197.

> Such actions – demystifying, denaturalizing, estranging – seem to offer important directives for highlighting the often invisible discriminatory functions of big data, but how exactly do they work? Who and where are the actual people implicated in big data discrimination?[296]

Presner positions the digital as being at odds with the ethical, seeing a human interface as a preferential and somehow more "ethical bias":

> are the 'digital' and the 'computational' at loggerheads with the ethical, and, if not, what might 'ethical' modes of computation look like in terms of digital interfaces, databases, and data visualizations?[297]

Indeed Presner's interpretation of the ethical with respect to the SHOAH Visual archive is itself problematic. He's querying the ethics of digital practice, but from my perspective it seems that the "dubious" ethical content is contributed by means of humanist intervention, and the computational is just getting the blame. Further still, he seems to be under the impression that the computational is somehow objective, when they are not, they are as vulnerable to human interference as analogue interpretative methods:

> Might, then, the realm of the 'digital' and the 'computational'—precisely because it is, by definition, dependent on algorithmic calculations, information processing, and discrete representations of data in digitized formats (such as numbers, letters, icons, and pixels)—present some kind of *limit* when it comes to responsible and ethical representations of the Holocaust?[298]

Presner's argument regarding computation ignores the inherent ethical implications already active (and underexplored) when it comes to algorithms and computer software:

> computation—as a genre of historical representation that includes data, databases, algorithmic processing, and information visualization—can be used against itself, so to speak, to not only deconstruct assumptions of objectivity and mathematical certainty but also give rise to a renewed attention to the ethical. As such, far from simply replicating the structures of automation and information processing used in the planning and execution of the Holocaust, I will argue that computation also contains the possibility of an ethics of the algorithm.[299]

Presner argues that datafication is de-humanising because the data visualisations in the SHOAH archive are built off of a double blind between form and content: "While computer-generated data visualisations may illuminate certain commonalities, patterns, or structures through quantitative analyses, ethical questions immediately come to the foreground."[300] According to Presner, this" abstracts and reduces the human complexity of the victims' lives

[296] Koen Leurs & Tamara Shepherd, "Datafication & Discrimination" ibid., 211.

[297] Presner, in Fogu, Claudio, Kansteiner, Wulf, and Presner, Todd, Probing the Ethics of Holocaust Culture.

[298] Presner, in ibid.

[299] Presner, in ibid.

[300] Presner, in ibid.

to quantized units and structured data. In a word, it appears to be de-humanizing."[301] And indeed in the context of the SHOAH archive he is correct to a certain degree, but Presner neglects to acknowledge the fact that these computational structures are already the product of human interference. He seems to perceive of computation as something that, because of a perceived objectivity, is somehow fundamentally unethical; but if ethical issues are introduced by the *human* programmers that encode the software, and the *human* contributors that compiled the keywords, does this computerised process not provide some sort of counter-functionality or anathema to this interventionist protocol? It's fine to argue that "computation also contains the possibility of an ethics of the algorithm,"[302] but to do so without dealing with the extant epistemological implications of these functions will lead to problems down the line.

Arguing for the "implementing [of] a practice of humanistic computing characterised by an ethics of the algorithm"[303] is fine, but Presner doesn't acknowledge a) the humanistic elements of computing as they are already and b) that the humanistic computing he identifies as active in the case of the SHOAH VHA is itself problematic. Presner emphasises the "flattening" effect brought about by facilitating access within a digital environment:

> Even as the Shoah Foundation's database assures factuality and facilities access and preservation, it has the side effect of flattening differences between the testimonies and rendering listening one directional.[304]

But surely (if this is not too pedantic a point to make) the same can be said of narrativised collations of testimonies? For example, does *American Slavery As It Is* not similarly flatten difference? The flattening effect of digitisation and entry into a database is one side-effect of the digitizing process, but Presner neglects the wider epistemological implications of software technique by arguing that they are "empty, neutral," when as we have seen this is not the case:

> A mode of organizing information characterised by the 'separation of content from material instantiation ... [such that] the content management at the source and consumption management at the terminus [are] double-blind to each other.' In other words, the content of the testimonies knows nothing of the information architecture, and the information architecture knows nothing of the testimonies. In this sense, the database is simply an empty, neutral bucket to put content in, and the goal of the information system is to transmit this content as noiselessly as possible to a receiver or listener.[305]

Presner argues that keywords manually entered by human project contributors, for example, are presented as an "ethical" alternative/ counterpoint to the perceived objectivity of computerised searches:

[301] Presner, in ibid.

[302] Presner, in ibid.

[303] Presner, in ibid.

[304] Presner, in ibid.

[305] Presner, in ibid.

> Without the visual interface, the database is still searchable by way of the tables containing structured data (name, place of birth, date of birth, date of death, family members, and so forth); however, the totality cannot be seen without an interface that visualizes the scope and scale of the database (which is —and this is critically important—a very different thing than the 'whole' of the event called 'the Holocaust').[306]

What Presner observes as "the impulse to quantify, modularize, distantiate, technify, and bureaucratize the subjective individuality of human experience"[307] does work both ways however, and the human imposition of "humanist" ethics themselves have ethical implications in terms of the elements they leave out. This keyword process delimits the archive not only to the keyword dataset, but also to the subjectivities of the figures responsible for entering the keyword material.

The compromises brought about as part of the data collection, creation, and inputting processes are widely acknowledged. Presner makes the well-known claim that datafication turns narrative into data whereas historians (and indeed researchers in general), create narratives from data:

> The effect [...] is to turn the narrative into data amenable to computational processing. Significantly, this process is exactly the opposite of what historians usually do, namely to create narratives from data by employing source material, evidence, and established facts into a narrative.[308]

Schäfer and van Es similarly note that datafication involves abstractionism:

> However, the translation of the social into data involves a process of abstraction that compels certain compromises to be made as the data are generated, collected, selected and analysed (Langlois et al. 2015).[309]

Edmond similarly acknowledges this but makes an important contribution in the form of reminding mindful and aware of "the concomitant processes developed not just by humanists, but by humans, to preserve and make meaning from these noisy signals."[310]

Computerised techniques and software packages have been identified as problematic from a humanities perspective because, as Rieder and Röhle note, they "wrestle" epistemological agency away from the human:

> While a critique of the standardization and streamlining of research through widely available software packages is important and raises many concerns, it does not tell

[306] Presner, in ibid.

[307] Presner, in ibid.

[308] Presner, in ibid.

[309] Mirko Tobias Schäfer and Karin van Es, *The Datafied Society. Studying Culture through Data*, 13.

[310] Jennifer Edmond, "Will Historians Ever Have Big Data?," in *Computational History and Data-Driven Humanities* (International Workshop on Computational History and Data-Driven Humanities, Springer, Cham, 2016), 96, doi:10.1007/978-3-319-46224-0_9.

> us how epistemological agency can be wrestled back from tools that make exceedingly complex methodological procedures available through simple graphical interfaces. A critique of digital tools is incomplete without a critique of their users and the wider settings they are embedded in.[311]

The following point will focus on error in big data research and bring up the issue of statistics, but statistics are also a good example of a widely used methodological approach that brings with it specific and difficult challenges that are not themselves under-addressed within the discipline of statistics itself, to say nothing of the myriad other disciplines that have incorporated statistical tools into their own research:

> Since the empirical social sciences have been using digital tools as integral part of their work for decades, applied statistics is a good place to start. One of the most widely used software packages in the Social Sciences is SPSS (formerly Statistical Package for the Social Sciences) and the significant reliance by researchers on this program begs the questions to what extent these scholars are capable of 'understanding' – or even seek to understand – the considerable methodological and epistemological choices and commitments made by the various analytical techniques provided.[312]

With the creation of digital software packages that facilitate the use of statistics, together with a growing mass of students and researchers that, as Rieder and Röhle note "are trained in using these tools without considerable attention being paid to the conceptual spaces they mobilize,"[313] we see a black boxing of more than just calculations:

> What is black boxed in such a tool is not merely a set of calculative procedures, which are, in the end, sufficiently well documented, but statistics as a field that has not only its own epistemological substance, but many internal debates, contradictions and divergences.[314]

> [...] many examples for the quite fundamental disagreements in the practice of applied statistics. While software can be designed in a way that highlights these divergences, it is too much to ask of a program to carry the weight of providing an education in the field it is mechanizing.[315]

2.2.8.iv. Error in big data research.

Moving on to error, concerns have been raised regarding the impact of error in big data research. Lipworth et al draw attention to the increased potentiality for false-positives

[311] Bernhard Rieder, Theo Röhle, "Digital Methods: From Challenges to Bildung" Mirko Tobias Schäfer and Karin van Es, *The Datafied Society. Studying Culture through Data*, 116.

[312] Bernhard Rieder, Theo Röhle, "Digital Methods: From Challenges to Bildung" ibid.

[313] Bernhard Rieder, Theo Röhle, "Digital Methods: From Challenges to Bildung" ibid., 117.

[314] Bernhard Rieder, Theo Röhle, "Digital Methods: From Challenges to Bildung" ibid.

[315] Bernhard Rieder, Theo Röhle, "Digital Methods: From Challenges to Bildung" ibid.

together with the misinformation that surrounds big data research which leads users (and researchers) to overestimate its capabilities:

> With respect to data analysis, big data leads many researchers to address a much larger number of questions, often without a clear hypothesis or with much lower prior odds than their top choices might have had. Thus, they are likely to operate in a space where the chances of getting false-positive, spurious results are very high (Ioannidis 2005b). Furthermore, many of the promises of big data, such as personalized (or precision) treatment, rely on extending concepts that have largely failed or have very high error rates, e.g. subgroup analyses which have long been known to have very low validity (Rothwell 2005).[316]

Of particular importance regarding the danger of false-positives is the fact that big data requires us to redefine our margin of error, because they are so large they defy traditional statistical rules: "With huge sample sizes, traditional statistical rules applied in other types of research may make little sense, e.g. p-values of 10-100 may still reflect false-positive results."[317] This is an example of the epistemological side-effects of big data research, which displays many of the drawbacks addressed earlier regarding regular data, just on a much larger scale; specifically in this instance we encounter the impossibility of producing discipline-specific data using discipline-specific techniques alone and also the implications of using these imported techniques. In the case of big data, traditional margins of error must be redefined which necessitates an awareness and strong working knowledge of statistics together with the facility to assess the wider epistemological implications of using such statistical analyses in relation to big data sets that would be far beyond the average humanities researcher (or indeed the average computer scientist).

The potential pitfalls of big data research have also received attention from Schäfer and van Es who note the tendency within academia to place “blind trust” in the digital:

> Although data sets can provide new insights that offer opportunities for fine-grained detail previously not available, their possibilities are frequently overestimated (e.g. Anderson 2008; Schmidt & Cohen 2014). Within academia, the blind trust in models, methods and data has been consistently criticized; recent big data enthusiasm has motivated a cohort of critical scholars to raise the alarm yet again (e.g. Couldry 2014; Gitelman 2013; boyd & Crawford 2011; Pasquale 2015).[318]

Schäfer and van Es acknowledge the work and contributions of Rob Kitchen in the form of his four fallacies:

> Rob Kitchin in The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences (2014) identifies four fallacies sustaining big data empiricism: 1. Big Data can capture the whole of a domain and provide full resolution; 2. there is no need for a priori theory, models or hypotheses; 3. data can speak for themselves free

[316] Lipworth et al., “Ethics and Epistemology in Big Data Research,” 5., 5-6.

[317] Ibid., 6.

[318] Mirko Tobias Schäfer and Karin van Es, The Datafied Society. Studying Culture through Data, 14–15.

> of human bias or framing; 4. meaning transcends context or domain-specific knowledge. (133-137)[319]

A knock-on effect of error in big data, is the loss of user trust. If, as Edmond argues, “scholars need to learn to trust the systems” then concordantly “we need to develop systems that support trust.”[320]

This stems in part from misconceptions regarding the potential of big data, but given the presence of a growing mistrust of big data among the science community, we can assume a certain percentage of this is transferred from users familiar with the problems being encountered in relation to big data in the sciences where we have concerns raised such as Lipworth et al’s regarding whether bigger is indeed better: "Overall the idea that big is necessarily better has very little epistemological support."[321] And also a palpable mistrust as to whether the so-called advantages of big data are advantages at all:

> Another concern is that big data may be promoted as being able to replace experimental, randomized studies for pivotal questions, e.g. documenting treatment effectiveness, with claimed advantages being more real-life settings, representativeness, and low cost. However, these advantages may be misconceptions.[322]

The growing mistrust towards big data is presented as positivist from the perspective of Rosenthal (albeit from a tentatively established connection regarding the importance and implications of big data visualisation and cognitive responses to visual stimuli), it is seen as a concern by Lipworth et al (2017) who observe "the deficiencies of observational studies (e.g. confounding by indication) do not get eliminated with big data, and in fact they may be compounded by the volume and often suboptimal quality of the information."[323] These hesitations are not only specific to scientific big data, as we can see from the following quote from “Narrative against Data in Victorian Novel”: "In its vast majority, big data research is observational rather than experimental."[324]

Among the science community there is a growing mistrust of big data. This is fuelled by miscommunications regarding the reach and potential of big data:

> A more general consequence of epistemological misconceptions about big data research is that the promises made by 'big data scientists' might simply not eventuate and big data research will prove to be (yet another) example of over-hyped biomedical science and technology. This over-hyping cannot be blamed solely on scientists, governments, and commercial sponsors of big data research, as it may also be a consequence of (un)critical attention of bioethics and other social sciences. While this might seem like a relatively benign issue, scientific hype is a substantial

[319] Ibid., 15.

[320] Edmond, “Will Historians Ever Have Big Data?,” 103.

[321] Lipworth et al., “Ethics and Epistemology in Big Data Research,” 6.

[322] Ibid.

[323] Ibid.

[324] Ibid.

> moral and political problem because it can undermine the overall credibility of science among the public (Caulfield 2004).[325]
>
> Many applications of big data research come closer to the public than other forms of biomedical research and may even involve the public as citizen scientists collecting and using their own data cloud. While this is exciting, failures of the approach will be more readily noted by the public and may generate further mistrust.[326]
>
> Of course, if big data initiatives fail to deliver on their promises, governments will also stand to lose both the trust of their citizens and the anticipated gains of biomedical innovation.[327]

In addition, the efficacy of big data in the sciences is being called into question:

> On the one hand, traditional research has typically had a problem of being underpowered to detect modest effect sizes that would be of scientific and/or clinical interest. This has been a pervasive problem in many biomedical fields as well as in many social sciences. In theory, big data can eliminate this problem, but they create additional challenges due to their overpowered analysis settings (as described above). Moreover, minor noise due to errors or low quality information can easily be translated into false signals (Khoury and Ioannidis 2014). Analysing rubbish or incommensurable datapoints may not yield useful inferences. The process of analysing many types of big data has even been called “noise discovery” (Ioannidis 2005a).[328]

3. Methodology

3.1 Interviews - list of interview questions provided in Annex 1.

3.1.1. Aims and Objectives.

The aims and objectives of the interview were as follows:

1. To obtain a more detailed picture of the various understandings of the term “data” that underlie computer science research and development.

2. To more fully understand where and how data is cleaned, transformed and processed, disassociated from context, imported into new contexts, and how ambiguity and uncertainty apropos data is dealt with throughout the research and development phases of a computer science project.

[325] Ibid., 7.
[326] Ibid.
[327] Ibid.
[328] Ibid., 6.

3. To obtain a more detailed picture of the understanding of the role played by narrative among computer science research and development.

3.1.2. Interview structure.

The interview was divided into five sections.

1. Positioning your Data Activities.
2. Assumptions and Definitions of Data.
3. Data cleaning, pre-processing, and Processing.
4. Messy data/ difficult material.
5. Relationship between data and narrative.

3.1.3. Participants/ recruitment method statement.

Participants were recruited from amongst the networks of the four KPLEX partners, specifically from TCD's ADAPT centre. We leveraged our association with TCD's ADAPT centre to recruit computer science researchers for the interview. The ADAPT centre's focus on personalisation and user content in particular gives them a downstream perspective on their work in terms of familiarity with user facing environments that will be of great benefit to the project. We were looking for looking for rich rather than large responses and so were aiming to recruit a minimum of 12 interview participants.

3.1.4. Interviews

WP2 generated a pool of questions that were then designed as a semi-structured interview adapted to our interviewees' scientific aims, and disciplinary background. The mean duration of an interview was one hour (ranging from 46 minutes to a maximum of 1 hour and 11 minutes). Researchers affiliated with TCD's ADAPT Center were contacted via email. The first of 13 interviews was conducted in July 2017, the last in September 2018. All thirteen interviews were conducted in a face-to-face situation. Following the first interview, the interview partner sent an unsolicited group email to all affiliated researchers in the ADAPT center "strongly urg[ing]" ADAPT Center scientists to participate in the interview and stressing the importance of the project. This email led to a marked increase in interest in the project and a near rapid influx offers to participate. The disciplinary backgrounds of our interview partners were computer scientists or computational linguists. It was challenging to find female computer scientists and male computational linguists, with the result that further emails were sent specifically requesting update from individuals that fit these categories. The gender of the interview participants was divided in women (7) and men (6). All of the interviews were recorded and then transcribed using the transcription software "otranscribe".

A total number of 13 interviews were transcribed, and anonymised. Anonymization involved the deletion of the personal names of our interview partners and their research associates.

3.2. Data Mining Exercise

The objective of the data mining exercise was to examine the occurrences of the word data and the surrounding text across a selection of major Big Data journals and conference proceedings; with an additional interest in examining terms relating to data cleaning, processing and scrubbing.

Our intention was to answer the following question:

1. Data & context. How big are the assumptions made around data the term?

2. Related question: Is this a problem? How widespread is it?

3. Are data transformation processes described? What terms are used to describe the cleaning/ scrubbing process?

The aim of this exercise was to help us understand the level of terminological instability and/ or terminological comorbidity that exists in contemporaneous Big Data journals and publications.

First, I consulted the Association for Computer Machinery Digital Library (ACM DL)[329] and in particular the *ACM database of Special Interest Groups* in order to compile a longlist of sixteen potential Big Data/ data publications that may be of use to the exercise. This list consisted of the following journals, all of which were available in either pdf or html format:

1. JDIQ Journal of Data and Information Quality

2. JEA Journal of Experimental Algorithmics

3. TALG ACM Transactions on Algorithms

4. TIST ACM Transactions on Intelligent Systems and Technology

5. TKDD ACM Transactions on Knowledge Discovery from Data

6. TMIS ACM Transactions on Management Information Systems

7. TOCS ACM Transactions on Computer Systems

8. TODS ACM Transactions on Database Systems (TODS)

9. TSAS ACM Transactions on Spatial Algorithms and Systems

10. ACM SIGDC Special Interest Group on Data Communication

[329] Contents of the ACM DL available here: http://dl.acm.org/contents_dl.cfm?coll=portal&dl=ACM&CFID=955676209&CFTOKEN=82355993

11. ACM SIGIR Special Interest Group on Information Retrieval
12. ACM SIGKDD Special Interest Group on Knowledge Discovery in Data
13. ACM SIGMOD: ACM Special Interest Group on Management of Data
14. VLDB: The VLDB Journal — The International Journal on Very Large Data Bases
15. ACM SIGMIS Database: the DATABASE for Advances in Information Systems
16. Journal of Big Data: https://journalofbigdata.springeropen.com/

Of these, the *Journal of Big Data* (JBD) was identified as the most appropriate journal for the purposes of the exercise.

Following this, and in line with our research aim of conducting a data mining exercise across both journals and conference proceedings, I identified the IEEE Transactions on Big Data,[330] the conference proceedings of the IEEE International Conference on Big Data, as the most as the most appropriate conference for the purposes of the exercise.

Given that both the IEEE Transactions on Big Data conference proceedings, and the Journal of Big Data articles were available in PDF format, and not in raw text, I first had to collect the documents and convert them from pdf to txt. The articles were packaged up on an article by article basis within the journal/ conference proceeding databases, which meant rather than convert one journal in one go, each article has to be converted one by one.

It was decided that the data mining exercise would only be conducted on articles labelled as "research." For example, of the 26 articles that appeared in *JBD* 2016, 21 categorised as "Research" articles. I randomly selected five research articles from each year, from both the JBD and the IEEE Transactions on Big Data. The rationale for selection is outlined below.

Rationale and process for the selection of articles from the Journal of Big Data:

Selected at random 5 research papers a year from the Journal of Big Data from 2014 to 2017 (Volumes 1-4). With the exception of 2014 as of the six articles in the issue only two papers were research, the others being a short report, two case studies, and a survey.

Each article was gathered (downloaded) and renamed so that the file name imparted full details of the journal, volume, author, article etc.

These were converted from pdf to txt using Cloud Convert (https://cloudconvert.com/).

Rationale and process for the selection of articles from the IEEE Transactions on Big Data:

Selected at random 5 conference presentations from the proceedings; preference was shown for lengthier proceedings.

[330] Available at: http://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=6687317

Each article from the IEEE database was gathered (downloaded) and renamed so that the file name imparted full details of the journal, volume, author, article etc.

These were then converted from pdf to txt using Cloud Convert (https://cloudconvert.com/)

Once all the articles were converted from pdf to txt files and merged the files by year, so that all articles/ proceedings from a given year were collected together into the one file.

I cleaned the txt files to remove journal metadata in the form of headers and footers containing the name of the journal, works cited, and references. The title of the paper (normally repeated at the bottom of each page) was removed with the exception of its first occurrence at the beginning of an article. The Journal of Big Data cleaning involved removing references at end of article, and the header/ footer “Journal of Big Data” or “J Big Data” from each page. The IEEE was not formatted to have the term data in header or footer, so this did not require cleaning in this way. However, in the case of the IEEE publications, transition from pdf to txt was not very clean: passages were rearranged, letters were merged or deleted, and the proceedings in general were littered with spelling and formatting mistakes. Because of the impossibility of getting clean base texts for the exercise, it was decided that the JBD files were sufficient for the purposes of the investigation.

Given that we were looking for content based iterations of the term “data” in descriptive or narrativised (as opposed to titular) environments, the recurrence of the title of the journal or article was considered unhelpful, and this is why it was removed. The works cited and references were removed for similar reasons: we were looking to analyse the narrative only, and iterations of the term “data” in the works cited would not have been helpful in this regard.

The individual txt files were imported into Voyant for analysis. A merged document containing the contents of all filed from the JBD was created, to give overall results and findings.

3.3 Data Analysis

3.3.1 Data Analysis of Interviews

The transcribed and anonymised interviews were coded and analysed with the help of the software Atlas.ti. A preliminary set of codes was created by referring to our list of interview questions. More generally applicable codes were added that not only focused on data definitions and complexity apropos data, but which could also provide potential links to the other work packages of the KPLEX project. These codes comprised e.g. “hidden data”, “uncertain data”, “complexity”, “context dependency.” Each code within the codelist was provided with a content description, its properties or brief examples for reference. These codes, most of them grouped within five code families, were applied deductively to the interview material, yet complemented by inductively emerging (in-vivo) codes. Ultimately, the code list comprised of altogether more than 45 codes (see Annex 2).

Some of our codes were later merged with other codes in order to create a reasonably and comprehensively applicable set of codes. Some of the codes on our preliminary codelist had not been applied and were therefore deleted. Ultimately, we obtained rich and thickly coded interview transcripts with more than 950 citations.

3.3.2. Data Analysis of Data Mining

The individual txt files were imported into Voyant for analysis. A merged document containing the contents of all filed from the JBD was created, to give overall results and findings. We specifically looked to identify the most frequent term used in each journal/ year: across all sample files this most frequent term was “data,” and the second most frequent term was frequently “big” as in “big data.” A merged table of results for iterations of the terms “data” and “big” respectively, spanning entries in Journal of Big Data 2014-2017, was created. These merged findings have been included in Annex 2 and Annex 3.

4. Findings and Results

4.1 Data Mining Results.

4.1.1. Journal of Big Data

i) The JBD 2014 corpus contained 11,572 total words and 2,063 unique word forms. The average words per sentence was 28.9. The most frequent words in the corpus were: data (134); resource (100); time (84); workload (73); map (70).

ii) The JBD 2015 corpus contained 46,603 total words and 4,705 unique word forms. The average words per sentence was 25.8. The most frequent words in the corpus were: data (1185); learning (249); big (225); deep (183); time (158). All but seven of the iterations of the word “big” related to “big data.”

iii) The JBD 2016 corpus contained 40,630 total words and 4,713 unique word forms. The average words per sentence was**:** 26.6. The most frequent words in the corpus were: data (930); big (329); stream (167); number (141); time (134). All iterations of the word “big” related to “big data.”

iv) The JBD 2017 corpus contained 50,021 total words and 5,483 unique word forms. The average words per sentence was**:** 24.8. The most frequent words in the corpus were: data (726); big (214); block (189); layer (170); storage (166). All iterations of the word “big” related to “big data.”

v) The merged JBD 2014-2017 corpus contained 148,826 total words and 9,751 unique word forms. The average words per sentence was**:** 25.9. The most frequent words in the corpus were: data (2975); big (782); time (537); number (368); using (353).

See Annex 3 for tables containing the Data Mining Results of the merged 2014-2017 corpus.

4.2 Results of the interviews.

4.2.1. The positioning of Data Activities.

Relevant Codes: Researcher Specifics, Interdisciplinary, Research Methods, Source of Data, Assumptions (about Data Source), Formulations of results and findings.

Researcher Specifics

The interviewees' areas of interest and specialisation spanned from "personalisation and information retrieval," "semantics and IR and then [...] personalization," design and innovation, data to text ("So taking sort of numerical or categorical data like spreadsheets and just creating a text summary or just description of the data"), "the approach of knowledge engineering techniques to management problems," and "Figurative Detection in Social Media" with a specific interest in sarcasm.

Two of the computer scientists interviewed had specific research interests in "the oncoming GDPR regulation," specifically in "privacy preserving personalisation" and

> how do you use that or how do you augment it to show compliance or how do you express everything related to it using semantic web technologies, which is this open standard of expressing data and knowledge. So, specifically I look at, how do you, like, understand the GDPR, how do you express it in a very computer sciencey sort of way, how do you define all the constraints or all the things that the GDPR mentions to do with data, to do with privacy and consent.

Four of the interviewees identified as computational linguists or as having a background and training in the fields of linguistics and computer sciences. These interviewees were more inclined to see themselves as "mediators" or intermediary figures, and to identify their work as interdisciplinary:

> I'm a computational linguist. So that means I have a background of computer science and linguistics.
>
> I'm a computational linguist, by training. So what I do is I work somewhere in the middle between linguistics and computer science, and what I do is I help normally engineers and computer scientists to understand the data, and I use their tools to process data, and then analyse the results from a qualitative point of view. And I work on Machine Translation mainly, so what I do is I focus my research on how people interact with Machine Translation, how professional translators interact with Machine Translation, and how end users interact with Machine Translation.
>
> I do machine translation and evaluation. Mostly human evaluation. But I'm interested in all, let's say, I'm a linguist by my graduation, my Undergrad was in linguistics, and then I had a Masters in computational linguistics and my PhD was in human

> evaluation of machine translations. So, I have one foot here in the School of Computing, and one foot there in the humanities.

In contrast only one computer scientist identified themselves as being in a mediating or "in between" position: "between computer science and statistics. So it's just like intersection between them." One interviewee identified themself "in the Digital Humanities space" and an emerging interdisciplinary research space that "involve[s] a bit of computing science in terms of human-computer interactions, kind of how humans interact with digital content and then it'll also bring in literature and media studies as well, in terms of narrative structure and how content is displayed online for users."

Interdisciplinary research

Interdisciplinary research recurred as a point of contention and potential conflict for some of the researchers, with differences of approach, differing standards, and a lack of respect of or understanding for other disciplines cited as having a negative impact on their research:

> The computer scientist doesn't care! They just need to have an agreed term, right? Linguists can be discussing that for hours!

As can be seen above, computer scientists' were presented by some interviewees as being indifferent to the nuances of the disciplines they were interacting with, or "not car[ing]":

> I always like to do a qualitative analysis of the results. What happens a lot of times in computational linguistics, or well, some people refer to it as natural language processing, is that you have computer scientists deploying fantastic systems and tools and applications, but they don't care about the linguistics.

This distrust towards the engineers and scientists they interacted with was particularly notable in relation to the computational linguists, who cited differing approaches to and standards for translation employed by computer scientists and engineers, approaches that favoured "metrics that take into account semantics and synonyms and things like that" over the knowledge of a trained linguist as a point of contention:

> What I struggle a lot on, especially with translation, is that they rely on a reference translation, a previous translation of the text that is being translated, and then they compute metrics against that. If you are a linguist you know that actually there are many other things that can affect the way you translate something, it's not only…and of course there are metrics that take into account semantics and synonyms and things like that, but still, a metric is a metric. It's a machine, knows nothing. It's just how the algorithm was implemented, right?

Aside from generating distrust and dissatisfaction with research methodologies across disciplines and the efficacy of the translation systems produced in such environments, interviewees with backgrounds in computer science also struggled with what they perceived to be an indifference towards reporting protocols on the pre-processing of data that would be considered standard or best practice in their field, but that were not being accounted for in

“Machine Translation papers” that were presumably written by people who did not have a background in computational linguistics:

> There are parts that are really taken for granted, and you will see that also in Machine Translation papers. It was really tough for me to get into the research area, because there are so many things in the pre-processing step. They just, many people writing their paper, that they do some pre-processing, but they never explain what. And I'm was thinking, But that's the key for me! How did you pre-process the text, right?

The role of the computational linguist was cited by one the interviewees as “putting back the linguistics in computational linguistics”:

> And the computational linguist is in between. It's trying to compromise and it's trying to make the pure linguist understand that the computer scientists need something else, and it's collaborating with the computer scientist to achieve a better result.

The interviewees with backgrounds in natural language processing or computational linguistics were themselves familiar with computational approaches and computer science methodology:

> I learned how to programme a little bit. I can do my own scripts, I can play with different implementations of algorithms. I know how to write programmes and the command line and things like that, but I am not supposed to create a new system. I will work with an engineer and I will explain to the engineer what I need, and I can read the code more or less and say, and understand what the code is doing. And discuss with the engineer why I need to change this part or that part because that's where I think the problem is. I was trained to be em...like a bridge between the pure linguist that only cares about the data, and the computer scientist that is programming.

They cited a lack of reciprocal familiarity among scientists regarding linguistics or translation:

> Cos, yeah, we have, you have linguists who are, people like me that have mainly a linguistics background. And then you have computer scientists that are working in that area but they don't have the linguistic background. They have the algorithms, the programming.

This criticism was not all one directional, with one of the computational linguists interviewed expressing bewilderment at how linguists with little or no training in the computational approached their research, and their ignorance towards the benefits of the computational:

> I've seen people doing all kinds of weird things manually in the dictionary and I think Jesus, if you had someone who know a little bit about code you could do this in a few hours. So, they use those facilities in quite a naive way.

Similarly, one interviewee who was particularly critical of computer scientists still acknowledged the need for computational techniques when it comes to analyzing data:

It's hard as a human to say Here's a load of data and then you can make a conclusion just from reading through a couple of hundred of them that Ah oh Irish is evolving and that People tend to use this and tend to use that, because you're making a grand statement just based on a few. So humans can't do that and that's why you need computers to find these patterns and to be able to give you the statistics

While there were multiple expressions of frustration regarding standards, best practice, and unfamiliarity with the discipline of linguistics expressed by computational linguists towards computer scientists, computer scientists in contrast saw their work as being resoundingly beneficial to the disciplines they collaborated with:

> So it's built by historians and then providing them with some tooling, so they could, you know, do some simple automated quality assessment on annotations that typically what happens is all the poor history PhD students do all the grunt annotation of primary sources and then you know, the more senior researchers have to sort of check it. They can't check everything, we've given them some tools to say Okay here's, you know, based on a few sort of rule of thumb checks, is where there might be some issues. So this is where you should focus your, focus your checks, you know.

In the case of translations and working with translators, their tools were considered to provide "hints" to professional translators regarding where and how problematic translations arise or helping historians annotate primary sources:

> we've done things like that in sort of for translators as well, sort of come up with some hints to say okay, this is where the problems might be.

> like say in helping the historians to improve the reliability of how they annotate primary sources.

While acknowledging that "we can't really help them exactly with their judgements, that's their domain knowledge," the computational expertise proffered by the scientists was presented as something that allowed them to "provide some tools to help them to catch some of their sort of omissions," while acknowledging that "the ideal is you know one of these techniques would automatically annotate those but I think we're an awful long way from doing that."

One interviewee referred to interdisciplinary research as something akin to mediating, a statement that an example cited previously wherein a computational linguist referred to themselves as "like a bridge between the pure linguist that only cares about the data, and the computer scientist that is programming":

> it's quite interesting to see how, you know, people in arts and literature approach narrative, and then how computing scientists approach narrative, because they have very different approaches, very different definitions of these two things. And so it's a matter of reading both sides of the story and trying to find a medium where both sides could understand what I would be talking about as, kind of, that middle ground. So, that's been interesting to see how they do that because the narrative

> people have, you know, they've done literature for years and they know it inside and out and then, the computing scientists are taking these technologies to be like, OK well we can use this technology to, like, change narrative and kind of mould it in a different way for the digital medium now.

This researcher cited the need to alter the vocabulary used when developing research or presenting research findings to different audiences, in other words, to audiences with different disciplinary backgrounds: “I think it would be up to the scholar, for example [...] to figure out which definition to use for which group when you're speaking to them.” Despite acknowledging this present day necessity, the interviewee expressed a desire for integration:

> So, while I might present the, you know, it to, do two different ways to do two different groups, I'd probably like to see some integration, and be like also this can be seen in this context too, to kind of give that awareness.

This desire for terminological integration can be counterpointed with the issues surrounding terminological comorbidity, lack of clarity regarding data cleaning and data pre-processing, and the overdetermined nature of the term “data” that will be discussed later.

The interdisciplinary nature of personal data was also cited as a major concern by the two interviewees with research interests in privacy and the GDPR. In particular the idea of “discipline specific consent” and “discipline specific data” was highlighted as a problem point that would come to the fore with the inception of the GDPR:

> what happens when you share data between different companies, because they do a lot, even without us understanding. Like someone might say Oh I'm getting all this data but I don't have the manpower, or I don't want to deal with analytics. So then they share the data with some other company who'll do that analytics for them and give the results back. But for the end user, how do I know that my data is safe with someone else. Under what obligations is that data being shared, how do you keep track of that? What data goes out, what data come back in? Is that under consent?

Research methods

Several researchers cited the acquisition of data as the first step in their research process, though for some this acquisition could only come after they had first identified the data: “my step for every kind of research on data is: identify the data, and getting access to the data.” Once this had happened, statistics on the data(set) were the next port-of-call “depending on which kind of work I’m doing on that, but it’s looking to some statistics about the dataset.”

Data identification, access and acquisition appeared to be the core facet of the research processes of the majority of interviewees, particularly those with a background in computer science, with one interviewee stating that all computer scientists are interested in [was] you know, what data can get hold off”:

> all computer scientists are interested in, you know, what data can get hold off. Our, you know, the sort of data we're looking for are often things like policies, work flows, procedures, you know, existing instructional, explicit instructions that organisations have established. Okay, so that's you know and then we go through a process of you know looking at that trying to interpret it, turn it into a machine readable model where we make some aspects of that knowledge explicit and we try and tease out what sort of a, you know, what sort of the common or recurring sort of logical concepts would be.

Identifying and gaining access to the data or "corpus" sometimes lead however to "see[ing] other things that you didn't think of. And then you need to come up with solutions." The data or dataset can influence the processes that follow after data acquisition.

Among the peoples interviewed, those with a background in computational linguists had the most methodical and clear-cut research processes of the groups interviewed:

> Well it starts being text in Spanish. Then we translate it with a Machine Translation System, and then what will happen is that we a new version of the same text in a different language. Then that text is translated, ah...sorry, it's post-edited, so corrected, so that's a third version. And then, and then you have the proofreader, which is the fourth version. So in the end you could even compare four different types, one is in one language and the other three are in the target language, right? But that, it's three different things. And then when you start annotating the data, and then trying to patterns and things then it gets even more transformed because that annotated data it's data in its own. Even the annotations could be part of it.

> first of all we need is parallel corpus. So we need the source language and the target language, whatever language is to be aligned in a sentence level. So, generally the data, like if you try to crawl like texts from the web, they are never aligned. So alignment is one of the things that we have to do, and cleaning the text, like getting rid of, normalisation, getting rid of accents and punctuation and all this kind of stuff. But I think the alignment is one of the most difficult things that we have to do, because if the alignment is not very well done, the translation is not good.

> But I generally organise my data as source, MT, post-editing, and reference, if I have a reference, for example, the reference would be if I have this text and I ask for a human translator to translate it, that would be my reference. And then I have, I use this in a machine translation system, and then I have the machine translation output, and then I ask a translator to look at the machine translation output and fix it, and then I have the PE, so I can organise like that. Sometimes, we don't have the human translated part, because there is no gold standard for the translation, so what I do is organise source, machine translation, PE, and my PE is also my gold standard because some human looked at it. And then I can have the source pre-processed and post-processed, or not and the PE, I can have post-edited the clean part, or also with the annotation for errors or not.

Aside from these, researchers working in the computer sciences also outlined clear and methodical research processes:

> So basically once I, there is now my hypothesis, my hypothesis OK: user's change interests. If I am able to track this change of interest then I can take into account this specific point in the user model to help the user to get better search results. So the first stage was looking at this data, the second stage is setting up a retrieval system that, given a query answers the user with some relevant web page, and my hypothesis is that, given a user that has a change of interest, given the user model, once I submit a query the ranking of the results change if I use the user model or if I don't, and my hypothesis is that if I use the user model then the results should be more similar to what the user expect. So should be more relevant what the user expect from the system. So, the second stage is now perform the evaluation with this retrieval system. Once this step is finished, I will present the results, usually there are some metrics that I can run on this, it's like especially for information retrieval there is a very standard way of doing evaluations and there are very specific metrics that you look and there are all in terms of in the ranked list of documents, where is placed the relevant document.

Others were fluent in the research process in and of itself, but nevertheless remained uncertain of how their research processes actually worked:

> all the machine learning, they're all just function approximators. So you've got your input data, you've got your output data, and you can write a programme to do it, or you can create a function approximator to do it, and it's just taking everything, representing it as vectors of numbers, and then using the tools that we already have in statistics, and all this kind of stuff, to approximate how you would map one set of numbers to another set of numbers in order to get the correct output. And it works. Somehow.

One interviewee spoke of the research process as one that was far less organised or methodical than the processes outlined by the computational linguists, referring instead to it as "brainstorming" and stressing the idiosyncratic nature of their projects as a reason for this less structured approach: "I don't think that it is a very kind of organised process it is usually just like brainstorming and asking questions, because I feel that it's in a lot of cases, very different." Despite professing to have a more relaxed approach, the interviewee nevertheless outlined a how a research project would develop:

> the first stage would be to get in touch with the industry partner and they would explain the problem they have, and from that we try to kind of understand what is the actual problem because sometimes the description is not really matching, and then we would ask some questions and follow up questions about like if this solution might help or something, and then we would usually have a meeting with the research staff, like with the people who have done a lot of research in the area that we think we want to solve.

The approach adopted, in the case of such industry directed projects, was decided on a case by case basis, and variables that influenced their approach included the persons or

personalities involved in the project, their knowledge of the area, and their backgrounds and areas of expertise:

> even the person. Like, sometimes the person really knows about the subject, they want to, like they would come with the problem but they would also have a deep understanding. But sometimes the people they would be coming from a marketing or sales background, so they know the problem but they cannot really know the technologies. So this kind of question it pretty much depends on like on case by case and the person himself.

Following this stage they produce a project plan:

> we would write, like once we have a consolidated plan or idea of what we're going to do, we would write like a project plan. It has like a description of the problem, and the proposed approaches, and who is going to work on it, who's going to do what. It's about like maybe 3 pages long, it's not very long.

While some interviewees (in particular, the computational linguists) were invested in their research project as it involved personally developed hypotheses ("my hypothesis, my hypothesis") or because it involved developing software that could be of benefit to at risk peoples ("because if the prediction are false or are not accurate, then what all of this for?"), others were less so, with the project plan and outline being seen as a "commercial document":

> Usually this kind of document is prepared by the project manager who is working with us or the head of the design and innovation lab. It's more like a, I don't know if it's like a kind of a commercial document or something, but it's more to show the industry partners that this is what we are going to do, and then they can sign it off, and commit that they are going to pay for this project or something.

These documents were referred to in another interview as: "a Business Canvass is basically like an A4 sheet and you have various columns, about ten or so, and the business canvas allows you to, kind of, put down all the points or it acts like a tool of discussion for start-ups, especially, so they can get all their plans and all stakeholders on one page."

Process and development was cited by some as trial and error:

> There would be, I think there should be like a proven process that says You should, like, it's better if you do this, but I don't think people follow that very strictly. They would just try the simplest thing, like at least for me this is what I would do. I try to do the simplest things that could work and then test that, and then move slowly towards like a better performance, like just try something on top of that and see if it improves the results or not.

But there was also the need to be open to the appearance of new variables:

> So, normally you start out with your background knowledge and your preconceived ideas of categories of content and that sort of thing. But normally through a pilot

> phase or a pilot test, you'll discover categories you didn't know exist for example, and then you add them to it. And then you have a pretty good matrix structure that you can work with, and then from there, like 99% of the data will be able to be categorised and catalogued and then you might find some uncertain data that you're, like, I'm not sure so this go in the other category.

And while openness to the appearance of new variables has already been cited as a potential outcome of identifying and acquiring the necessary data, one interviewee professed to being unclear as to how the process they were engaging in actually worked:

> I'm not a linguist. I mean you could say…but even the maths behind it isn't really theoretical. It just works. In fact, it's kind of a point of contention that we don't really know why it works, just that it works really well.
>
> And it works. Somehow.

Data cleaning and the pre-processing of data was cited by one interviewee as an integral part of the research process:

> what I would do to normalise all the corpus I was using beforehand—and I explained that in the papers—and sometimes people were asking Why are you saying this? And I said Well this is important because it changes whether the computer thinks it is one word or two words. If I have the same word spelt in two different ways but I know as a human that that's the same word, it's just two conventions of writing it. A computer would still treat them differently, so I have to unify that. And I think that's important.

This will be covered further later in section 4.2.3. Data cleaning, pre-processing, and processing, but it is important to note that while the cleaning and pre-processing of data was discussed by all interviewees, few mentioned it when asked to outline their research process.

The methodologies, tools, softwares and research processes employed and adopted by the interviewees were diverse and encompassed:

i) Machine Translation that can come with "a gold standard or gold set of reference translations":

> "machine translation"; "machine translation output"; "different types of Machine Translation systems. The rule based system is using a grammar and a dictionary, and it would be like trying to teach the computer and code everything that is in your brain".

ii) Annotating data and tagging, which is sometimes done formally, sometimes informally ("So, you know so sometimes the annotation has been a bit informal, we're trying to get a bit better at actually publishing annotations"), which could be annotated on "a spreadsheet" or by using a "more sophisticated system like EnVivo":

> “A small set that you start annotating, and you use it as a pilot and then you edit it, find out the issues that you have to solve. And once you have those issues recorded you can enhance the guidelines for annotators, or you can change the taxonomy if that's a problem. Or yeah, you have to correct the things that you are encountering because there will always be things.”; “So now I am working on this dataset, and basically I am using the words that the user use for tagging the page like a query, and then I have the content of the page.”

iii) “tools developed in the natural language processing community [that] use raw text.”

iv) “using rule based things and deep learning [ing] to extract information from data.”

v) “automated metrics for natural language generation.”

vi) using an array of software and methodologies such as an “an encoder or a thing called a multi-layer perceptron.” Using “RDF triples,” using “vector of numbers,” using “language model decoders.”

vii) “machine learning. You know, so you basically gathering lots of data and try and find correlations, you know, using various statistical neural network techniques and you know, we’re, that’s very diverse, like this technology is being used all over the place, you know, so they’re you know like one of the really big areas is how it’s being used to target advertising for Google and Facebook, so that’s very topical at the moment.”

viii) “parallel corpus”

ix) “teams of annotators;” crowdsourcing: “I also get, I call them “crowd,” so I work with crowdsourcing a lot”; “Mechanical Turk.”

x) “human-computer interactions”

xi) “kind of, draw[ing] from each discipline what I need to, kind of, mash them up together to create kind of a framework that I'll need to move forward in my research.”

xii) “Google Analytics”

xiii) “content analysis”

xiv) “semantic web technologies, which is this open standard of expressing data and knowledge”

xv) data “expressed in ontologies and linked open data formats.”

xvi) “SWRL rules.”

xvii) “Using Haenszel-Mantel or something, statistical formula.”

xviii) Using “biasing network[s]” with “joint probabilities, marginal probabilities, and conditional probabilities”

xix) Using “Application Programming Interfaces,” the adaptability of which dictated whether it was useful for a project:

> “the adaptability of the API, whether it can be coded by Java or C# or C++ because there are lots of bias in network modulars like huge in Winbox, many others but they are not like, they cannot be modified or cannot be adapted to your system developer.”

xx) Using “natural language processing that works with a syntactic parser.”

xxi) Using a “data dictionary”:

> I think, a data dictionary would just be like, yeah, if you, I mean you probably don’t wanna know this sorta stuff but if you’ve got, like say, a categorical thing and there’s like three or four different categories in that column, then you want to know what each of those categories matches up with because people, I mean, when I name my variables I try and make them descriptive, as much as possible, but some people just put like a single letter and it’s like meaningless.

xxii) Using software for anonymising: “you have various technical means of anonymization such as KLMT, t.Anonymity and so on.” The use of this software however raised issued for this particular interviewee, who observed that another researcher could define personal data information differently, which would directly influence the material removed:

> So, someone else's definition of what can be considered as personally identifiable may not be as extreme as someone else, in which case they have a problem if they're working on the same set of data. So, if both of them are saying OK we are going to remove all identifiable pieces of information from this data, one person's dataset is going to be much larger than the other person's. So, sometimes when, like in that case, you're literally changing the definition of what is data. Like, this data may be a part of some other dataset, but how I derive this data from the bigger dataset itself defines what the smaller subset is. So that's, like, how do you define this relationship between different pieces of data? That needs to be clear. So, that's where you have gaps in communication.

One of the most commonly cited factors that influenced the early stage of a research process was personal connections, such as having family members or friends with specific professions, or having particular language skills:

> “So I contact my friend, Doctor in Indonesia and I spoke with epidemiologists in Indonesia”; “So now it’s restricted because you have to start with something, and it’s doctors because I have relatives who are doctors. It was easier for me to get participants”; “I am Italian. So we choose that one for obviously reasons it was the

> language more near to me, and we performed the evaluation only on the Italian language."

While many interviewees had set methods and processes for conducting their research, one interviewee cited a preference for exploring "as much as possible" and experimenting with newly developed tools and software "new technologies new projects all the time": "I like to try new things and just I'd be able to explore as much as possible." This same researcher, whose work proceeded by trial and error was, not surprisingly, comfortable with making mistakes and having to begin again:

> I think you get a chance to try more things and you have the freedom to do that, because even if you, if something doesn't work or didn't work as expected, it's still OK. It's part of the research.

This relative freedom was cited by the interviewee as a "luxury" of "computer scientists of this era":

> I think like computer scientists of this era at least, they have a luxury to say like It's not only that I want a job they could say Ok I want a job but I want a job that gives me freedom and where I could try my own ideas, and they could get that because it's just there are many opportunities at this point. But I think that like in any domain like you would like to have maybe 10% that you want to kind of try new things or do something that you really like. Sometimes like people cannot do that unfortunately.

When in such an environment where experimentation and innovation are not only encouraged but facilitated and incorporated into the working day and the development of research projects, the decision to opt for one technology instead of the other was cited as a personal preference, technique or expertise: "Em, there are many and it's just, you know, your personal preference but guided by experience that make you choose some specific. More than programs, it's like which kind of technique you are going to use."

Any given project has the potential to receive more data input or add "different dimensions of data," with this being augmented if the researcher has access to more tools or, as above, has access to newly developed software:

> you could even have keystrokes, like recording exactly what they did in the keyboard. And if you have an eye tracker like we do in ——— you could even record where their eyes are going, to identify the words that are more problematic. So then you have like different dimensions of data. There is text but then there you have the time, and you have the fixations of the eyes, and all that together has to be merged somehow to, to answer your research questions.

Source of data

Interviewees sourced their data from a wide variety of outlets, expressing awareness that data can be of varying quality and provenance, with the quality often directly linked to where

and how it was sourced, collected, or scraped. Interviewees cited material “scraped directly from the internet,” or text copied from a website:

> There's different kinds of material or data that I use. So, some of them could be something I directly take from a website. Literally just copy text, a bunch of text.

For others, this “scraped” content was domain specific or even terminologically or phrasally specific (Google ngrams). For example, as in the case below, the dataset was specific either to the domain of Facebook or Twitter:

> For us it will be the collection of the corpus that we call as a data and then we choose the domain of the data, is it from coming from Facebook or is it coming from Twitter which area we want to choose and sometimes the collection of data is also different in all the resources available for that domain like Twitter.

One interviewee was basing their research around the “Amazon review data set,” while another sourced their data from datasets released by large companies such as AOL: “a first small dataset that was released in 2006 by AOL. It is about query and click information [...] I think it’s about one month activity.”

One interviewee expressed awareness that this “stuff that has been made available online for researchers” had, to a certain degree, already “kind of not exactly” been exposed to processing and curation:

> I’ve mostly been using kind of not exactly pre-processed data but stuff that has been made available online for researchers.

Much of the data sourced had already been structured, or even “very structured,” though with the exception of the individual who observed that the data they accessed was “kind of not exactly pre-processed data,” this does not seem to be a point deliberated over by the researchers:

> most of that data I use is very structured because it's expressed in ontologies and linked open data formats. So, you have a very structured way of keeping the data. But then you should know how, which data contains what exactly. So, then you would have ways of annotating it. So, you would have a bunch of files somewhere that actually contain the data but then you'd still have to keep record of why I have it, why I'm using this and so on. If I have data sets then I would usually keep them in a database, if I want to query them. If I want to publish them, then you don't publish the database, then you publish it in an open format such as a spreadsheet or a data dump;

> translation data, so when professional translators or non-professional translators translate a document, they tend to break it up into sentences and translate sentence by sentence. So that gets collected as data which has from traditionally been reused as if you’re professional translator, you look at, you’re translating the latest version of Microsoft Office manual for Microsoft and you’d have, they’d give you a database of all the translations from the last time, the last version was translated

Whereas the above data—translation data, structured data "expressed in ontologies and linked open data formats"—appeared relatively straightforward and monovalent in terms of content, data sources can also be multivalent, and contain input from a variety of sources, containing both signal and "noise":

> it's a mixture of data from the web, from literature, from newsfeed, from...there's some legal text in there. So there's a whole array of different types of text that makes up that corpus. I also did some work on analysing tweets for Irish. So that's social media content, and that's referred to as user–generated content and it's noisy data so it was a little bit of a different angle.

> What kind of the data we have actually, we have the data about the concept, how can I explain about that… In our model, we have the data about the different concept, for example, as I mentioned about the user and some demographical data about the user, some financial records, some health records, some educational records about the user. Then we have some data about the policy to have accessing this data. For example, some policy comes from yourself, for example, your consent. Some policy comes from the government, actually, and we have the data about the policy as well. For example, which data in which situation, who can access this data? And, what kind of access that he can, actually. Then we have the data about the request as well, and what is the request, and who is the owner of this request, and which kind of data? The other things that we model is about the context, it means that about the environment that you are in, for example. The requestor in context, for example, who is the requestor, or what is the role of the requestor? What period of time he is requesting, and where is the location that he's sitting on? And we have the context about the data owner as well. Who is the, Where is the data owner and what time he is, for example, she's happy for sharing the data, and blah, blah, blah, things. And we have the data about the kind of access that we can give to the requestor, for example, access to, type of the access actually. Read off the data or write off the data, or modifying the data, deleting the data. What it is the, which kind of action that he can do. And the other kind of data that I have modelled…action, requester, I try to remember these things.

The content of datasets are often not as straightforward or giving as the researcher may wish them to be, with private corporations such as Google or Microsoft refraining from releasing data in its entirety, perhaps because doing so would compromise the data, or because doing so would give their competitors an advantage:

> So most of the literature on this topic is about, is from people that actually work in Microsoft, Bing, or Yahoo, and Google. So they just release their metrics about their methodologies applied on those dataset, but they don't release the dataset itself.

Another interviewee cited the "dark data" as a source of data that was particularly challenging to work with and decipher, being almost entirely without what they referred to as "usable context":

> It takes a lot of effort to clean it and to put it into a usable context whereas this dark data stuff, it's just …

> I: What is dark data?
>
> R: It's like… Alright, so you have a spreadsheet of data but you don't have what any of the stuff in it means, so you don't know what the columns mean, you don't know what the column titles are, you don't know what the contents mean, you know, or you have, you do know what it means, but there's like hundreds of different possible things and you're not really sure which corresponds with which description. So, it's just, it's you have a lot of data but there are so many different possible definitions or ways to describe it, that you can't actually use it. So it's kind of, it's almost unusable and there's a lot of that and there's some companies are working on it.

The source of data can be project specific, as can the factors (practical or otherwise) that influence why a researcher may choose to work with data scraped from that particular source or environment:

> For my work, sarcasm, I choose on Twitter because it was social media. I choose Twitter for two reasons. First it is small text, easier to crawl, and there are many other resources already has been used and there is hash tag (#) which was first introduced which can help us do the self-annotation so that's why

Data sourced can also be knowingly incorrect, or contain a margin of error, particularly when it comes to crowdsourcing as a means of data gathering:

> Well, in dataset collection when you do crowdsourcing in this case, since we have to depend on the author annotations so the one challenge we had is we had to trust him. Even if we see that it's not at all sarcastic also there is a lot of noise we have. People like stuff, even if they don't mean it. If they like and then they unlike, that will create can create us a problem, because whenever you like it's going to be automatically stored. So it won't be, machines cannot understand when it is liked and it's changed back. So that's one trouble we have with the like system, we can have it. Also the complex reply which is not expected, which was not expected at all. [...] So those sort of answers is also not expected, are completely unexpected, and we have to process those also.

In addition to errors of mis-identification, listed above, data quality can also be compromised by the researcher responsible for collecting or curating it, and data quality can also vary when it comes to long term projects, such as for example the digitizing of long term linguistics corpora:

> I suppose em, talking about the dictionary then, that consists of twenty-three volumes made over a period of seventy years. People that died and then someone else took over. The standard differs hugely from volume to volume. I mean it's now digitised but it reflects, it's merely a copy, an annotated copy of the original, and so the quality differs from every volume to volume. You don't see the volumes anymore now but letters so A was done quite different from P. So em, some information is not included for some words. And then others it's, the info that are, would be quite handy for me

> isn't included in some part, and in others it is or sometimes it says A late form. Now if it is, if it deals with a thousand years what is late and what is early? I don't know.

One associated problem with digitization projects that seek to replicate the analogue is that "It doesn't lend itself to structure or to structuring" and may be particularly difficult to parse:

> From a modern point of view, from a data capturing or digitization point of view it's very incomplete. It doesn't lend itself to structure or to structuring sometimes.

Linguistic data appeared most often to often take the form not just of "raw text" or "raw machine translation," but of textual corpora:

> A collection of texts [...] digitised, digitised texts. But not annotated in any way, just raw text;
>
> the dictionaries weren't compiled from the electronic corpus, the paper dictionary already existed, it was scanned, and smartly marked up so that it's easier to search it, but it's basically just a copy of the analogue thing. But it's easier to search through it, em, I mean that's not, that's, that work, a large part is a manual endeavour. You scan it and then you have to take the errors out, some xml, some formatting might be automatical, but then you have to check it. My approach, building a morphological analyser is something that, that does that automatically;
>
> in a lot of cases it, em, you only really, the real, really only eh, resources for language are just lexical resources which are electronic copies of dictionaries;
>
> The main would be in UCC, a corpus of electronic texts CELT. That's the main one and it has, I don't know, how many words does it have, I don't know. But it deals with other languages as well, but, so I would deal first of all with already existing texts, with already existing digitised texts.

One interviewee, with a background in linguistics and extensive experience in working with machine translation, availed of data in the form of subtitles:

> the OPUS corpus and the corpus of Ted talks, as well. So, all of this were available online, right, so they got all that corpus, they trained the machine translation system, and all the outputs, we work with 11 languages, so the output for all 11 languages, what I get from them. So, that's my data;
>
> data from the MOOC platforms that we have, like for example, Coursera and Iversity. So they provided us transcript of the subtitles and the, all the data that they have in the forum and, you know, the comments and everything;
>
> now we have Opus corpus so I just go there and just take whatever they have texts from the web.

Data quality can vary depending on the source:

> noisy data of like Facebook or Twitter or Youtube, I know that there's going to be a lot of bad data there like the grammar would be bad, there would be a lot of misspelling. So I would try to kind of to sort those issues, but if I get like some kind of like the Wall Street Journal data I would maybe assume that it's more or less very good.

Whereas in the discipline of computer scientists, data was not as likely to be purely linguistic or text based:

> all computer scientists are interested in, you know, what data can get hold off. Our, you know, the sort of data we're looking for are often things like policies, work flows, procedures, you know, existing instructional, explicit instructions that organisations have established. Okay, so that's you know and then we go through a process of you know looking at that trying to interpret it, turn it into a machine readable model where we make some aspects of that knowledge explicit and we try and tease out what sort of a, you know, what sort of the common or recurring sort of logical concepts would be.

> ultimately we're doing is to try and come up with some explicit sort of machine readable classifications of knowledge so there, so we use existing languages. So for the description logic there's a set of standards from the World Wide Web Consortium who are, so they're sort of quite well established and that's interesting because that standardised and is starting to fit into a lot of database products now as well. So that allows us to sort of have classifications and extensions to particular sort of knowledge vocabularies and then the other area, yes, so that come up with what would be known as a sort of open data vocabulary so we would do a structure of that information and that would then allow us to annotate, we would use that classification when we're annotating you know, a particular piece of text or particular sort of piece of source material, let's say you know for GDPR it could be the law itself, it could be these other sources that have already been annotated, we annotate their annotations. We're also looking at things like you know, we look around for other sources and the ideal is to work with the company and get access to their information but they often will list data protection, extremely cagey. You know, because they're extremely, they're worried about researchers coming in and spotting the big hole that could result in prosecution.

Sometimes the source of data is a pdf file that has been annotated or made "machine readable," and this annotated dataset, a processed or altered version of the original document, becomes the source, as it is easier to work with than the original pdf document:

> like the source data would be you know written content typically. You know, often we're pulling things off the web, sometimes it's a sort of a webpage, sometimes it's a sort of a can be a PDF document. When we are trying to sort of really annotate that closely, you know, our preferred process is to sort of do our own sort of clean canonical version of that. Canonical is not the right word, but a clean version of that because you know, if you haven't seen HTML, it's a load of messiness that you can have in HTML

> I was looking at the actual legal text of the GDPR and it is online, which is good, because these days, it's like everything is online. People can say Oh here is the url for something and you can go and access it. But then it's in a form that you cannot refer to someone, I cannot tell you Oh go look up Article 18.2 whatever and you, being a human, know that you have to visit that url, scroll down or find Article 18.2 and see whatever is written there, but how, for a machine this is not trivial. You can't just say Oh do this the same as a human would do. So, then I had to look at how do I transform that legal piece of text into some way that machines can understand. So, then I had to break the individual articles, points, sub-points, give them each an identifier. So, in essence, it is the same data but now I have added additional annotations which will allow you to point towards different bits of article. So sometimes you get data which is ok and fine and what is the standard way of representing it, but you can't use it in your research, so you have to annotate it additionally in order to make it usable.

Others may be extensively cleaned versions of a data source:

> one example is the Data Protection Office, so the Government regulator publishes all their decisions. So that's also a good source of information. So, you know the result would be, you know, we would you know, have those documents. What we usually try and do is publish a like a clean version of those documents.

Other examples of reports or textual documents (pdf or otherwise) availed of by the interviewees include reports, or statistics/ ratios derived from "raw data" such as WHO reports, and UN data.

Variable levels of awareness regarding the implications, ethical, privacy related, or otherwise (for example, validity of research) of sourcing data and the need for accountability re the data source. Two interviewees were particularly eloquent when it came to outlining the need for clarity regarding "the purposes of the data. And the research. And the source. And the curation of it":

> Where did the data come from, what form was it in, what did it involve, who gave it, do you have a historical record showing this? So that is provenance information so that's, like, the first and foremost challenge is how do you put down this provenance information?

One interviewee, versed in the idiosyncracies of the GDPR, made clear that when speaking about data source, one must also include metadata, and that this marginal or peripheral material could itself provide personal information about an individual:

> So for example, you have data that is being stored on let's say Dropbox or Google Drive, and in terms of how Google Drive or Dropbox sees it, it doesn't care what the data is, right? According to the legal definitions, it's just a storage service. Like, you can store stuff on it but then it doesn't know what you're storing in it, so that's how they get away legally being responsible for it. But then they still store who is storing it, who are they sharing it with, what kind of files do they have, is it videos, is it text files, Word files, is it titled CV? So, all of this metadata that is involved with the data, if

> tomorrow they start using this kind of data, then that's an obvious change in how they're seeing data. Then how do I represent, or how do I even face, or how do I deal with this change in data in my research?

For some, the sourcing data was considered unproblematic:

> these days as you know that we don't have any problem to finding the data because of the, today's technology, for example of the cloud computing, and something like that, we don't have problem with finding the data.

Another interviewee, similarly noted the relative ease with which they can sometimes access a desired data source, with the data being provided, or "dumped" by an industry partner:

> we would just get the data basically dumped by like a partner or someone who wants work on it.

For others, the sourcing of data may occur as the result of serendipity or because of disciplinary or cultural backgrounds:

> *at first I have no guidance but after talk with a physician, epidemiologist, so like they give me some inputs. They give me some materials like the valuable materials like Atlas of human infectious diseases and they give me like WHO reports, annual reports about environment and climate change. So I can observe the weather and location also.*

Sometimes the source of the data is a core facet of the project in and of itself and identifying it can in fact be the aim of the project:

> Sometimes part of the solving the problem is to find the data. So for the, I told you the offensive content detection. It was part of it was kind of to be able to find offensive material that's already tagged. So the partner didn't have that. We had to look for sources. So we were able to find some things and then first challenge was to kind of be able to get what you need from all this data, because they would be on a website with a lot of clutter around. So you want to take for example specific part of the text. So part of the process is to extract these interesting pieces, and then be able to store them somewhere, and make that efficient enough, because if it's a very large dataset, it could take long time to process it. Try to kind of make it as quick as possible, and then on the data itself like, for example, working with text, textual data there is some pre-processing that we would do, depending also on the problem. It's not always the same, but sometimes we might decide to remove the punctuation for example or just lowercase everything, or like remove the characters that are weird and like for example some data comes from Twitter and has a lot of strange characters, or emojis, or like people sometimes write, like they would write one word and then repeat the same letter like I don't know 10 times as like emphasising or something.

Data source can itself be subject to change over time, meaning that once scraped, the "source" may undergo modifications, with the scraped data no longer resembling the source exactly:

> If one of the companies that I'm saying OK this is my example company, I'm going to check their privacy policy, I'm going to find all the uses of data that they do, and then I'm going to put it in a paper and say Oh we did our use case over this company and then they change their privacy policy, and people are going to be, like, It's not up to date. So that's a very, not the kind of data you would expect someone to deal with but that's also change in data. Someone changed their privacy policy my research is based on and now my research is kind of out of date, which means that it uses outdated data, even though traditionally people may not think of this as data. So, actually you have various definitions of data that people implicitly assume, but they just don't term it as data.

Other data, once sourced, required extensive cleaning in order to be used effectively:

> So what we would have, just the core text and just a core set of tags inside it. So we would use the standard HTML tags and then other, any other ID tags we try and control which ones sort of stay in there and which ones we strip out. So that's partly just to make it a little bit easier to manage the annotation of it but also we need to, you know, the reason we do that as well is because we can't rely on the source data staying the same. So if we just point to a, to a you know a really big long document, you know like a legal text on a website, depending who it is but we, even if it's all very reliable like the EU, they may still do an update and you know, do a lot of changes. You know, what you would read on the screen might be the same but there might be a few like highlighted errata, but the actual underlying HTML could be a rejigged quite a lot, you know, so for example often in HTML you will have essentially bookmarks, you know, you like tag a bit of text and you will say this has an ID and the ID typically comes from a database or something. It's not you know, it's automatically generated so, that's something that you can refer to. Say I can, I want to refer to that particular article or that particular clause but if the document gets, if you like, re-indexed automatically and it's only being done for publications, so nobody else is annotating it, then they don't really care that the numbers are different. They just care that they're unique, okay, but if we've separately annotated and put into a number that they've generated that, that throws us off.

Despite the very best efforts of a researcher, their research may quickly become based on "outdated data" if the material they have based their research on is updated, with the result that their research may come with the caveat "We got this data until the period of…and then you specifically mention...":

> So, if tomorrow you're using a bunch of privacy policies and someone updates them, then you have to say that OK, this is no longer valid, I'm going to create a new one with the new privacy policy and then, rerun my work over it. So, at some point you have to, like, stop and say that I cannot incorporate new changes anymore because I have deadlines and so on. But, ideally you would just keep on incorporating changes

> as much as possible. But then that brings in the question that are you aware what kind of changes happen over time?

Little distinction was made between “raw data,” “pre-processed data,” “processed data” as a data source, with the data present at the onset of a project, rather like the NASA EOS DIS data; being considered as “raw data”:

> “Like Indonesia, Kampala, Uganda, and Bissau. They have like their own ORs and they are all different and then I have to transform those ORs into like pool ORs so I can get, I can get this code OR as their transformations.”
>
> Like OR OR is, no, OR is like summarised ratio from raw data but I treat them like data for my research.”

Two of the interviewees, neither of whom had a background in linguistics, were either currently working on, or had experience of working on projects whose data was in a language they could not speak, or understand:

> I: Can you speak German?
> R: I don't. So that is part of the challenge, that we need to kind of have a tool that is not relative with a specific language.
>
> You know like, sometimes the barriers, we have that come up quite a lot is just language, you know. So we work with a lot of multilingual content and when we’re undertaking a project, we have to go around and say Okay do we have anybody who can speak Spanish? you know. Then if we don’t actually you know, like you can parse it you know. Often we may be doing something which is just parsing it and formatting it. But we still we think need a native language speaker because we can’t check if we’ve done things wrong, you know, then we can actually look at the, the you know the structuring of it and just make sure we haven’t done something stupid whereas with an understanding you can go Well no you’ve put that in as a heading and it’s plainly a footnote, you know, or something like that.

Other sources of data for the interviewees, particular those working with linguistic data, are corpora of data specific to a previous research project released and available for subsequent research projects. Release of and access to this data by third parties is contingent on obtaining consent from the participants:

> So, what we ask the participants of the experiments to give us the data so that we can release it for research purposes only, as a corpus. That's one type of data that we have in computational linguistics.

That said, several of the interviewees expressed awareness that data acquired for use within one project, or for a specific purpose, could be used or exploited by others for other purposes. Therefore having repercussions beyond the sphere and remit of the original research project; which again brings us to the issues of ethics and data privacy issue, and specifically to the GDPR:

> Once you have access to data, like even, you're recording this, right? You could even use this interview, if you were working in speech recognition. You could use it do a speech recognition of non-native speakers of English, and then train the system. So this interview could be data in itself, for someone with a Spanish accent, because I know I have a Spanish accent! Everyone, once you have data you always have different things that you can use.

> the other thing I think that people don't really understand is what happens when my data, the question is always Are you happy for your data to be used in a certain way and you sort say well what's the worst that can happen? Okay, but it's something that's way more dangerous combined with hundreds of thousands of other people's data, you know, what could happen in that case, you know. So you may not give away a lot of information about your health condition for a particular app, but you give away a bunch of other information about how often you walk, what you eat ,or something like that, and then somebody else can correlate that against similar data against people who do have, for who you do have health conditions. Then you can start making inferences about your health. Where you know, you certainly didn't think you were handing that over. But how do you explain that to someone, you know, easily? So perhaps some sort of you know, narrative about why would somebody want to do that, what is their intent? Are they, you know, what sort of virtues are they bringing, or what virtues they're ignoring in making their decisions.

Open data, while it brings privacy concerns such as those identified above, was identified as a key step against the deprecation of data:

> some of these other organisations we talked to, they may just deprecate the data, you know, they may not have an archiving mechanism. So that becomes you know quite important to us and also the places where we publish put emphasis on that because they're all about, you know they're also in the game of providing open data.

Assumptions about data source

Of the assumptions articulated by interviewees about the source of ones data, the most salient (and perhaps alarming) were directly data quality, whether or not the data was "accurate" and whether or not the data is "real." These concerns bring to mind Rosenthal's definition of data as a "fiction," but whereas Rosenthal was specifically referring to data used in fictional environments to grant plausibility to an otherwise fictional narrative, in the case of one of our interviews, access to "real" data was a concern as otherwise the research itself would be based on "not real data" or "simulated" data:

> OK, all this research that is done with user basically should involve the user itself, but most of the time that's not possible like in this case, so basically you are assuming that the data that you have are kind of reliable in terms of simulating the way that a real user would interrogate or question a real web search. So this is a strong assumption, most of the time it's not true. So, basically whatever conclusion you get you should always frame in the right context like, most of the time what you would do

> is looking at very specific behaviour, so specific that you can actually discover them, even in this kind of not real data.
> [...]
>
> I: Yeah, and by not real data you mean like?
>
> R: Simulated, like a simulation, and the best practice would be like confirm this kind of accounts that you have got in a real scenario, so involving real user making like an AB test, these kind of things.

Examples of “real data” cited by the interviewees included up to date user behaviour, with older data sets that may once have been “real” or have reflected accurate browsing patterns, for example, subsequently devaluing with time and, as a result, losing their status as “real,” with “real” here seemily implying current or up to date data. Other examples cited were health records and medical data:

> Real data. Let me tell, if I bring a sample maybe it will be much tangible. For example, imagine that I am a patient and normally I use the hospital X for my periodic check-up. It means that I have some health record in the hospital.
> [...]
> Okay, and the hospital knows about this kind of, the health record of me, this is the hospital X, okay? And, there is some policy for using the data. For example, I gave a consent for using my health record data when I went there or I will go there for doing some check-up, okay? For example, imagine that one day, I’m a diabetic one, diabetic patient, and I went in to visit my GP to do some check-up, as a normal monthly check-up. And when I visit my GP, normally he ask about some process and some data and now after these things he need to add some data to my health record. And, as I gave the consent to using my health record for the hospital, and since the GP is one of the employer of hospital, then GP can access to my health record, okay?

One interviewee highlighted the privacy concerns of working with “real” or “true” data, specifically because the sharing of this data becomes problematic:

> But the other problems comes on actually, for example, finding real data or which one is a true one or even, privacy comes as one of the problems with big data or actually in sharing the data, okay?

Incorrect or misinformed assumptions about the data source can lead to incorrect conclusions being drawn about the data, and indeed the tools used to analyse the data:

> So if you’re not aware of the source of your data or the nature of your data, and then you’re going using tools on this data and then you could conclude going Oh this tool is no good…

Ironically, given that Rosenberg’s definition of data as “rhetorical” was a prominent part of the literature review and early to mid-research phases of the project, access to data post-GDPR appears to have the potential to hinge on rhetorical manipulations. The lengthy

quotation that follows outlines a scenario wherein consent surrounding data privacy and data access are inferred rather than explicitly sought:

> And in that Regulation, they mention that you can't use personal data without having any consent from the data owner, explicit consent. But, my research don't go strictly for that one, but we are looking at the similar things. I say that, okay, you are completely right, we have a cons… the patient gave a consent to using my health record data for the GP that is inside the hospital, okay? But, there is some situation, other contexts that we, we should to, know about the GP. For example, GP should be inside the hospital to having access to those data. For example he can't access to my health record from home. Or even sometimes you can give this consent as well because sometimes some, in some emergency ones, maybe the GP needs to have access from home, but it depends on your consent actually. And, but I am saying that the GDPR says that, Okay, we will, we must to have some explicit consent to using this data. I am saying that, Okay, what will happens for you if the other, for example, your nurse comes to helps you in the emergency one and there isn't any explicit consent that shows that that nurse can have access to your data, okay? The GDPR says that No you can't have access to those data. But I am saying that if you say in a high level that, for example, any healthcare giver can access to this data, it means that the nurse can also have access this data, okay? Where this explicit data, explicit consent will come? Actually from our modelling the data. If we can model the data in a hierarchical level, for example, I say that I have a model that shows that the nurse and GP are employer or are members of healthcare giver. And, as, for example, G—— is a nurse, then based on this model and that consent then G--——— can access to this data, okay? Our main purpose is inferring this kind of privacy policy from the existing one. And some part is completely explicit and we can infer directly, and some part not explicit. There is some implicitly one from this model in the data, and even sometimes using the semantic web technology we can infer and bring some reasoning on those data to inferring the data. If we can't infer these kind of consent, then we contact with the user, it means that data owner, and ask him or her about the accessing this data. We say that, Okay for example, G--——— needs to have access to your data and there isn't any implicit or explicit and we can infer that she can access to your data. The purpose of her access is for this kind of things. We model those data as well, for example, for example, as a one concept in our system is request. We should to model the request to show that who, what data you requested, who is the owner of this data, and for what purpose do you requested this one? It means that we model all of these things and based off this model, we can say that, okay, for example, G——— needs to have access to this data because of these things, and are you happy with sharing with those data or not?

Two interviewees outlined scenarios wherein the assumptions that guide data sourcing are knowingly compromised or inaccurate:

> I'm just taking for granted that what somebody who published or annotated or edited the text, that's what matches my input. Maybe he made mistakes, he wrote, he misread words in the manuscript. But that's, I'm not concerned with I that. I just take the digitized version.

> Well, in dataset collection when you do crowdsourcing in this case, since we have to depend on the author annotations so the one challenge we had is we had to trust him. Even if we see that it's not at all sarcastic also there is a lot of noise we have. People like stuff, even if they don't mean it. If they like and then they unlike, that will create can create us a problem, because whenever you like it's going to be automatically stored. So it won't be, machines cannot understand when it is liked and it's changed back. So that's one trouble we have with the like system, we can have it. Also the complex reply which is not expected, which was not expected at all. [...] So those sort of answers is also not expected, are completely unexpected, and we have to process those also. So that's [...] kind of strays a pretty bit, an upstream task, can we use this one also to find out what is the senti… if that was sarcastic or not, even if the reply was sarcastic or not. So it is like in inception like, you are going again and again. So, we haven't gone through them, we just want to build the system this one first and then we see what happens next.

In both cases, awareness or suspicion of the presence of inaccuracy or error did not result in any modifications to the data.

Formulation of results and findings

Results and findings were made available in two ways: through publication, and through the release of data or datasets.

> The basic reason for publication is like there's two things, one of them is to either present or allow someone else to reference, cite, or use, and the other one is validation through peer review. So, those are the two basic ones all researchers would go for.

How data is represented in publications:

> In a traditional paper, you would have mostly text, some diagrams to kind of give an outline of whatever you're describing. So, in my case, if I've an ontology it would be like a bunch of graph notes, maybe a table describing something, but like, these days, it's often encouraged to add in additional stuff. Like if you have an attractive visualisation of anything, then you can't put that in the paper, but then it's encouraged to put in a url saying, oh you can access this here. And then you can go online and actually play with it online

Best practice in terms of the formulation of results and findings is to "publish your annotation on the web" while adhering to community standards for data representation:

> you come up with an ontology, because you sort of need to back up, you know, why is that good one, so that the best practice is to actually publish your annotation on the web as well and get their you know, there're standards for doing that that we adhere to.

It is not enough, however, to make something available on the web, these annotations must be persistent, and being associated with or attached to an institution helps with this:

> So if we refer to the annotation from our publications, the annotations can be persistent on the web now. That's, there are techniques for doing that. The, you know, part of the problem is you know we don't really have a institutional, American institutional platform for doing that. So ultimately, for any sort of prominent [22.34] also sort of persistent URLs, you know, so your web address, you need that to be persistent. There is ways of doing that on the web, but you know, to make that really reliable you need to have an institution behind you

Persistent URLs also help safeguard against the deprecation of data, as does the use and provision of open data:

> some of these other organisations we talked to, they may just deprecate the data, you know, they may not have an archiving mechanism. So that becomes you know quite important to us and also the places where we publish put emphasis on that because they're all about, you know they're also in the game of providing open data.

Elsewhere, when working in newer media formats such as Twitter where a change in username may lead to the loss of data, one interviewee outlined how the tweets relevant to their project were made persistent through the creation of a Twitter bot and the retweeting of material relative to their project:

> So what we did we created a new account and we re-tweeted all of the 8,000 tweets and we shared the ID and we didn't delete that. In that way we actually can get around their system. [...] Yes, we just make it available. We, actually it's a Twitter bot so the data is there. So what people have to do is follow our bot.

Peer validation was cited as a key motivation for publication, together with the idea that cumulative release of data allows results in the development of "this big sort of network of you know models that are published on the Web" with a view to generating what one interviewee referred to as "some sort of canonical knowledge about certain modules":

> you know, because of the web, the whole idea is you should publish them and make them for other people to see and then other people can look at them and try and see are these the same concepts, are they different, and it builds up this big sort of network of you know models that are published on the Web, and right in the middle is to kick-start it if you like, is what's known as Dbpedia, which is a, which is all what's referred to as the structured data that's been gathered out of Wikipedia because Wikipedia has been a big global effort to try and have some sort of canonical knowledge about certain modules but it allows people to discuss it and try and come to some sort of agreement.

Data was represented in tiered or layered formats, in the form of primary data, annotations and published findings:

> sometimes our primary data could be something like a legal text. What we're doing is publishing that, publishing our annotation having our conclusions in the paper. So we publish those as data.

A single document may be stored in multiple different formats, such as HTML, PDF, and TXT, with the result that these data sources would be combined into a so-called "canonical version":

> We look at the official sort of HTML document from the EU, which is you know, the actual reference for somebody making a judgement on the law and they will have it, usually they'll have a text version, a PDF version and a HTML version, and that's then I'll canonical version underneath generating all of those, and you look at the HTML version and it's sort of done as a table okay, so it's got all the sort of clause numbers and then the clause there which is logically from a parsing point of view, you know, it implicitly has the information there but a table isn't the right construct for representing, you know, a set of clauses and a set of some bullets and things like that. And all those materials are available in the HTML, you know, they just don't haven't them properly.

The most common modes adopted for the representation of data include text files, spreadsheets, :

> Currently, it's just text files, mostly text files, I mean yeah. I mean there's also other stuff as well like, kind of like matrices, definitely matrices, and some spreadsheets and stuff.

Other modes of representation included tabular presentations, with the content of the tables being informed by the aim of the research:

> Usually it depends on the aim of my research. So for example like looking for the suitable base in network modular, I will make a tabular content of research. Then it depends on my research whether I have to seek the commonness or the adaptability with Java or Java environment or maybe the Python or maybe the OR, so those tabular will be like the items that I have to look for when reading those articles.

> We showed the tables because we tested on six data sets. Our data set is one and other five publically available data sets, or the system is available. We tested on them and we showed a table. We showed, it's actually the paper will be also available in SLweb right now.

Results and findings, and the representation of data associated with research and ones results and findings are often structured, expressed in ontologies or LOD formats. alternatively, if the database is not published it may be released in "an open format such as a spreadsheet or a data dump":

> most of that data I use is very structured because it's expressed in ontologies and linked open data formats. So, you have a very structured way of keeping the data.

> But then you should know how, which data contains what exactly. So, then you would have ways of annotating it. So, you would have a bunch of files somewhere that actually contain the data but then you'd still have to keep record of why I have it, why I'm using this and so on. If I have data sets then I would usually keep them in a database, if I want to query them. If I want to publish them, then you don't publish the database, then you publish it in an open format such as a spreadsheet or a data dump.

In addition to being represented in a structural manner, two interviewees recounted their experiences of working with semantic web technologies—specifically RDF and Portage—to structure or serialise their data.

> I use the semantic web to structure my data. And, because I'm planning to use machine to understand my data, and because of that I use the semantic technology for those structures, like I said.
>
> RDF is [...] a standard to serialising the data.
>
> I am using Portage as a tools that bring or creating concept and relation between the concept and individual for modelling my ontology.

These tools establish relationships between data, with one interviewee justifying their use of RDF by arguing that "just the data own self doesn't show any things":

> Because the, just the data own self doesn't show any things. The data relationship bringing some meaningful for us, actually. And the semantic web advantage is modelling the semantic knowledge, actually, and, yes, for example, as I mentioned, the RDF is one standard to putting that. It's really simple one, that just is so that you should to explain everything in three part.

Semantic web technology, and the structuring of data was presented as a necessary process in order to make the data machine readable, to establish meaningful relationships between data (both within and without a given dataset or database), and to "explain" the data. Once structured within an RDF, the data are considered "more understandable for the computer":

> RDF or Resource Description Framework is one standardised for showing the data. And what is the structure for this one, that is much more understandable for the computer one. And it shows that what is the subject and what is the property and what is the object.

The need for data to to be structured was mentioned elsewhere as a necessity by one interviewee who outlined the need for their material to be rendered isomorphic:

> There are several parameters that I use. For biasing network, I use the term isomorphic, so the knowledge base will have the similar structure with the created biasing network and I can see it by seeing, by observing whether all the nodes are

> created, whether all steps are created and if those are accomplished then it is isomorphic and the second one is by reading one by one the combination created in these, the condition of probability tables, and I have two measurements, it is precise and correct or accurate.

Elsewhere, one interviewee with a background in computational linguistics who was particularly reluctant to use the term “data,” nevertheless talked about the need to divide up “the predictable and the unpredictable, and with the unpredictable.”

When representing data, one interviewee noted the importance of excising any material that cannot be effectively or meaningfully structured, such as “uncertain data”:

> So, as much as possible you move towards structured information. Like, uncertain data is chaotic data, you don't know what's in it. So, you want to get rid of that uncertainty, you want everything structured proper, everything states something reliably. So, ideally you want to get rid of the uncertainty as much as possible.

In the above interview, the interviewee twice referred to data as “structured information.” The implications of this terminological comorbidity will be discussed later in this document.

How data are represented varies depending on the what the researcher “wants to do with it” (“I: What factors influence how you set up your different data structures? R: What I want to do with it.”), it depends “the data itself and the project,” with certain structures or representations being adopted “just [as] an ease of use efficiency kind of thing,” and finally with the intended/ envisioned audience in mind. One interviewee, a digital humanist working between computer science and cultural heritage commenting that “I feel like I'm going to have to present it in a quantifiable way to show them numbers or graphs or something that they are used to seeing”:

> because I'm an arts person and I'm working with computing science, so I feel like I'm really going to have to, you know, come up with a creative way to present my materials as data to the computing science crowd.
>
> So, while I might present the, you know, it to, do two different ways to do two different groups, I'd probably like to see some integration, and be like also this can be seen in this context too, to kind of give that awareness of the Digital Humanities, you know, cross-discipline aspect of it.

Of particular interest was this same interviewees stated reluctance to use the word “data” when speaking to researchers from the arts and humanities and citing this resistance among arts and humanities researchers to seeing cultural heritage materials as data, and a comparable bemusement among computer science researchers in terms of similarly identifying cultural heritage materials that are particularly difficult to datafy as data. This interviewee cited this as one of the particular “challenges of working with, you know, cultural heritage materials and saying, like, this is data” and opted instead to use the phrase “mode of content” instead of “types of data”:

> the way you would talk about these kinds of materials with, you know, the arts folks, it would just be, like, these types of things, for example, the types of mode of content is probably how you would speak about with arts, but then you would, with the computing scientists, you might say, like, OK there's these types of data and then there's this much of each of it, and this is what it means, like, so you'd talk about it more on a, like um, you know, how much, you know, that kind of level rather than what it is.

The representation of data, even in its abstract form, what one interviewee described as “the concept of data” is of particular interest to those working on the specifities of data privacy, issues surrounding consent and the GDPR:

> I'm not actually dealing with data per se, but I have to deal with the concept of data. I have to deal with how would you legally specify data, how do you represent like an abstract object called data.

In the case of translation, representations of data are often made using parallel corpora, parsing programmes, or tree banks:

> I think, a data dictionary would just be like, yeah, if you, I mean you probably don’t wanna know this sorta stuff but if you’ve got, like say, a categorical thing and there’s like three or four different categories in that column, then you want to know what each of those categories matches up with because people, I mean, when I name my variables I try and make them descriptive, as much as possible, but some people just put like a single letter and it’s like meaningless.

> so basically I developed a parser, so it’s a tool that will take an Irish sentence and analyse the syntactic structure and say That’s the subject and That’s the object and That verb and that preposition go together and That noun and that adjective go together. And then that feeds other language processing tools downstream so it can improve machine translation, or information retrieval systems, or language learning systems. So grammar checkers and stuff like that.

> And then this value is saying, This is the head of this word, and this, This word modifies this other word. So I’m gonna actually…I’ll get you a picture of a tree to show you what it looks like [...] So this is the root. This is saying This is the nominative subject of the root. That’s an auxiliary of the verb. And then this is what’s called an open compliment. So it means that they share the same subject so X compliment, open compliment, this whole thing is. This is an infinitive marker for this verb here, and so on. So that’s how it’s broken up, and the way you know how to attach them is these numbers that I was showing you. And so what that’s saying is that the direction of the arrow is towards the dependent, or the modifier. So that word is modifying this. And this word is modifying this. And they’re called head independents. So that’s what this information is here.

The third and final quotation above is noteworthy for an instance of terminological comorbidity wherein individual datum, when analysed or parsed, are subsequently seen to contain “information”: “So that’s what this information is here.”

However, these well-established structural arrangements are not always viable, depending again on the aims of the project, such as in subtitling projects, for example:

> They cut down sometimes three, four times, because subtitles have a specific time that it has to be shown on the screen.
>
> So, the alignment of that is a very hard thing to do, you know, so you have to, it’s loads of processing and trying, if you try to align by the time it shows up, because you would assume that it has to show up at exactly the same time because of the time the person is speaking, but sometimes it’s not, sometimes it’s like a few, one second before and one second after because it’s longer, so you know.

As noted in the section dealing with data sources, at times, the received data may itself not be represented in a usable manner, or in a way that facilitates interaction, with the result that such datasets may be annotated and re-represented so as to increase usability and interoperability:

> So sometimes you get data which is ok and fine and what is the standard way of representing it, but you can't use it in your research, so you have to annotate it additionally in order to make it usable.
>
> So, you know the result would be, you know, we would you know, have those documents. What we usually try and do is publish a like a clean version of those documents.

Sometimes the received formatting of data is, as noted below, “not like a normal html page any sane human would write,” with the result that the received data must be modified to make it usable and interactive. Such re-representing of data was seen as “kind of normal”:

> when I was trying to get the, when I was trying to annotate the text of the GDPR, it was not like a normal html page any sane human would write. It was actually like the entire page was made up of tables. So, each point was in a different table, and you'd kind of just copy and paste that and annotate it. So, you had to first convert that table text into normal text and break it down by point and then, do it, like, do the annotations. So, it's all weird kinds of stuff like this. Which is kind of normal.

When pressed for the reasons why this representation of data was considered particularly difficult, the interviewee elaborated

> It's in a non-standard format, or it's easy accessible but the way the way it is structured doesn't make it helpful, like I cannot use that structure, or it's missing some things that I need to be there, or I have to combine different datasets together to get what I want.

Non-standardised data representation impedes data usage, interoperability, and the comparability of different data(sets). In addition to these non-standardised formats, as outlined above, the viability of data representation is affected by data sparsity or missing data.

Another interviewee noted that not all data can be standardised because, when it comes to certain events (they cited an historical event), it is impossible for people to agree on the data:

> Sometimes people can't agree, so controversial issues don't get agreed and get marked as things haven't got a consensus around it. You also see for example that you know it doesn't try and like Wikipedia doesn't do much translation okay, they say well they do it in certain like physical sciences where there is an objective agreement on, you know, maths or anatomy or something like that but loads of other things they say well no, different countries and cultures should find their own version of Gallipoli or you know, World War II or something like that because they will have their own, you know, genuine perspectives on it you know and even within a single language, people are going to disagree about it and the process of writing a Wikipedia page is almost a process of deciding what you're going to leave out because you can't agree on it.

Other factors that contribute to incompatibility across datasets are different coding systems which result in what following interviewee refers to as "a problem of representation"

> you know, in computer science you have these different coding system, and if you choose the wrong one it's like you are not representing the data in the same way that is comparable.

While the representation and explication of all available data was acknowledged as best practice, or the "the ethical thing," as one interviewee put it, not all people or companies "want to reveal the secret ingredient of the recipe!":

> So many people just, ah...sometimes it's because they don't want to reveal the secret ingredient of the recipe! Sometimes it's, if it's a company participating or writing a research paper, sometimes they don't want to tell what exactly they are doing because they, that would give a competitive advantage to their competitors. So there are different issues that arise there. From my point of view, the ethical thing would be to actually explain everything.

Representation of results and findings are also researcher dependent, for example one interviewee described themself as visual, and her representations of data were visual as a result:

> It really depends. So, for example, it really depends, it depends what kind of correlation I want to do. So, for example, if I want to just show the numbers of in percentages of a type of error, for example, I find out that neural machine translation is making loads of, I don't know, word order errors? So, to compare that with the

> other system, the statistical system, I can use just like, you know, I can use a graph or I can use, like, a pie chart, or whatever, you know? But, if I want to show all the numbers that I have, like here are the numbers of errors in the presentation, the standard deviation and everything and then I'll do it in a table. But I do everything. I use all the visuals. I'm a very visual person, so I think that it's easier for me to explain my data with visuals as well.

Industry dependent, representation contingent on the project:

> we would write, like once we have a consolidated plan or idea of what we're going to do, we would write like a project plan. It has like a description of the problem, and the proposed approaches, and who is going to work on it, who's going to do what. It's about like maybe 3 pages long, it's not very long.

Not just in computer science, but also in computational linguistics:

> And it was funny how then you subconsciously just because of the type of client you have, you choose different ways. And then we came up with two different perfectly valid translations that were hinting towards one or the other direction because when you were at the situation where you had to choose, you chose the one that were, that was given by your client. You were thinking of the client that was receiving the translation. When you don't have that instruction then it will be your mind. You will be your own bias.

Role of error in the representation of data, and the formulation of results and findings:

> I mean, especially in language stuff, you would get kind of ambiguous sentences, for sure. I mean, I'd be getting too detailed if I was to say, like you know, it's stupid, you can get obviously you can get like one star reviews probably on Amazon, where someone says something good like Oh I liked everything about it but the delivery was shit so… one star, but in terms of like what I would consider to be data, with like spreadsheets and stuff, like…I can think of an example there: sometimes the decimal place might be wrong. So, a column is represented in thousands whereas it should be represented in millions or it should just have the actual itself rather than just being like a short hand or something, and then you can actually have different, you know, one column is a million, one column is a thousand and if you have that then it's ambiguous. But that's, it can be ambiguous if it's just not well specified.

Just as in the previous section it was noted that sometimes sourcing data was the project aim in and of itself, for others, finding a way to represent the data in a meaningful manner was, according to one interviewee, the "the purpose of the data itself":

> sometimes that's the purpose of the data itself, when you're given a data set and said what is the meaning, you know, what's the average, or what are the patterns? So you can , you know, you just, you kind of print out distributions and stuff like you know, what's the average, what does the, plot all the numbers on it graph and see how they spread out. You can look at correlation plots and stuff like that.

4.2.2. Assumptions and Definitions of Data.

Data definitions

A number of interviewees—all computer scientists—expressed a reluctance or hesitancy towards defining data:

> I'm pretty sure that there is a very specific definition in computer science at least but I don't remember it!
>
> I don't think like my opinion is that important.
>
> I don't have a perfect text definition of data.

Two expressed hostility or annoyance towards the idea of there being multiple definitions for data, stating in contrast that "Data are data!," that "the building blocks they are very simple, there is nothing. Data is just data like." Similarly, they felt that data was self-evident, "it is a thing" and "just exists" and is "just data":

> I feel like it's too self-evident. It just sort of exists.
>
> I would say by comparison that data exists, it does exist, it just exists in of itself.
>
> I would say that it is a thing, like I don't know about like attaching a meaning like that to it. It's just sort of, it is. I mean, for people's like sanity, and like ease of work, it's certainly makes sense to create data in the first place that is structured and make sense, but I don't think you can say like, it's not really anthropomorphic, but it's definitely like, I don't know what you'd call it when you call something good or bad. But it's not like, it's just data.

Others elaborated on how hard it was to define data, nothing on the one hand that while "everything that is not very well defined can cause confusion," it is nonetheless very difficult to define data:

> I think everything that is not very well defined can cause confusion, but at the same time I think it's a very hard to define, to have like a single definition of data.
>
> it's always very hard because in the research world, if you don't define exactly what you mean, but what you're saying, you're going to have problems and then you got, and then maybe that's what people means that data is not truth or something like that. But, yeah, it can cause, it can cause misunderstanding, so what I think needs to be done, because we still don't have like a single definition of data, is that researchers, they need to define exactly what is that data, where is it from, how was it handled, and how was it created and, you know all the steps that are needed to be done

Despite the difficulties associated with defining data, “What is your data” is nonetheless a question researchers are regularly asked:

> I run into this problem all the time because my project is supposed to be data to text but everyone is like what’s the data that you’re using.

Elsewhere, others were less reluctant or reticent when is came to elaborating on their working definitions of data, with one interviewee arguing that data are facts that can be numerical or textual:

> I would like to define data like facts, collected through facts. So maybe refer to my undergraduate and my Masters studies, it is like numerical facts that given in periodical time and, yes. It is, sometimes it is not always numerical but also in a sentence or alphabetical. Yes. I define it as facts.

Another defined data as truth, but as truth that can be modified by the researcher:

> Because for me, data is like the truth, all the truth are there but it is our way to modify or to summarise those data that makes them looks bad or good or fit or not fit to our research.

For another interviewee, data is figurative and mercurial, possessing the ability to mean different things to different people:

> here also right now data can be many different ways like as you said, multiple forms it can be, it is really figurative like people have different. So, for us data will be the collection of, so a data set we call it, so I would like to know what you mean by here data for different people. Is it the definition of the data, means if it is for me sarcasm what is the definition of sarcasm, or what I call data?

Multiple interviewees argued that data can be anything and everything “in the real life,” “everything you saw,” “what I am collecting right now,” everything is or has the potential to be of use to the researcher “to study a certain subject” or “prove a point, or not prove a point”**:**

> We can say everything is, in the real life is data.
>
> I liked these definitions of, like, at the beginning that you showed me with data, and it’s like all of those, all the things that it is.
>
> That’s all we think about, but other than that, right now, yeah, for me data is what I am collecting right now is an annotated corpus.
>
> Data for me is everything that I can use to study a certain subject, right?
>
> for me data is everything. It’s what you need to, what you need to study a subject and to maybe prove a point, or not prove a point. You know, it doesn’t matter really, the result, it’s just like you need to study. And it can be anything, it can be like a text, it

> could be interviews, it could be, I don't know, speech. It could be anything, whatever you need to do the thing. That's my definition of data!
>
> You know that the data is everything that you saw. Then every day, how can I say, I bring some, for example, data, some sample for showing that, for example blue is data, pen is data. But it doesn't have any meaning. Information, meaning the, means that, the meaning of the data. Where your data has meaning, it means that now they are valuable for you.

As one interviewee outlined, data is evidence, and its truth value depends on the questions being asked of it, and the individual doing the questioning:

> I think it is all completely spot on because I you know that whole question of you know, can you extract truth from data. Like, data is just evidence. I mean the truth is completely based on who you are talking to, you know, what the question being asked is and then you know, these will propagate through, you know, exactly what, you know have you selected and selected that data. So that's, I think, absolutely the you know crucial part of what you're, you know, what your sort of data analytics and knowledge engineering is being able to really show what the providence of our processing of that data, so what did we select.

For another, data is "just any material that you have in hand," but "digital material" in particular:

> I think data that is just any material that you have in hand. I think of it as like digital material. Anything images, texts, everything. It doesn't matter what it is, but anything. It could make, you can benefit from it, make use of it, or not. But it's still, I think, any kind of digital material is more or less data.

Many of those interviewed presented very insightful responses to my questioning them on the nature of data. As the following excerpt makes clear, certain people have very specific perceptions of what data are, but for others, and particularly in a digital environment, "literally anything could be data" to the point where this becomes "implicit data, you don't actually provide it but it's just given in any case":

> Some people's perception of data is, like, they have a very specific view of what data means. So, if they're using a website, they think data would mean only something that they enter or a product that they put in the card. But from the marketing, marketers, or the shopping website's point of view, literally anything could be data. Just the fact that you have access to the site is data in itself and which is also represented in legal text. Which is why you get that little notice saying "We use cookies." So that's because even before you've done anything, you already have submitted some of your data, like your IP address, what the browser you're using. Companies use this kind of information to create a personal profile for you. So that's another kind of data. So, there's implicit data, you don't actually provide it but it's just given in any case. But that is still your data. So, data in any case is information, whether you give it explicitly or not.

Data can therefore be “anything that you obtain from the user”:

> So, these are very difficult to define terms and even legal texts does not have clear applicable definitions of it. So, then you kind of have a working understanding of what data actually means. So, from my perspective, as an assumption, you just basically call data as anything that you obtain from the user.
>
> even before you've done anything, you already have submitted some of your data, like your IP address, what the browser you're using.
>
> typically once it is digitised and becomes a computer science issue and the distinction between content and data is that content is a form of data that is either produced or consumed or both by people. You know, so it’s got a language, linguistic or communication element for people, whereas, you know, there is lots of forms of data like the sequence you know what the temperature is in this room ever hour. You know, that would be a sequence of data that isn’t necessarily ever gonna be read by a human being, it doesn’t need to be.

This can be problematic when it comes to sensitive material such as medical data or personally identifiable information because just like that which is personally identifiable is open to debate, what one considers data (and specifically what one considers to be usable data or data that has been given consensually) can vary from person to person, from company to company, and from engineer to engineer:

> So, someone else's definition of what can be considered as personally identifiable may not be as extreme as someone else, in which case they have a problem if they're working on the same set of data. So, if both of them are saying OK we are going to remove all identifiable pieces of information from this data, one person's dataset is going to be much larger than the other person's. So, sometimes when, like in that case, you're literally changing the definition of what is data. Like, this data may be a part of some other dataset, but how I derive this data from the bigger dataset itself defines what the smaller subset is. So that's, like, how do you define this relationship between different pieces of data? That needs to be clear. So, that's where you have gaps in communication. That's why you have legal text that should very clearly and unambiguously mention what it is and so on.
>
> So, do you know exactly what kind of data it is? So, if someone says that Oh, this is not personally identifiable information. But someone else comes along and says No, wait a minute, this bit, like this sub-slice of it is actually personally identifiable, then that's kind of, like, a clash in information. So, it's not possible for someone like me to make a system that will automatically say, Oh this is personal information, this is not personal information. You can do that up to a certain point, for example it's easy to say Oh this looks like a name or This looks like an address. But then what happens when someone gives out like a random name that's not in your system? Then you don't know whether it's a name. So it's, the basic assumption is that whoever is declaring that knowledge knows what kind of data it is, knows what kind of metadata is involved, and that they keep it consistently in the system, they have this data consistent in the system. It's not like, today this bit is personal information, tomorrow

> that bit is not personal information. Then the whole thing kind of breaks down because you're actually changing what data means in that context.

A recurring point among those interviewed was that data is anything they are analysing. This is a subset of the “data are everything” responses, with the “everything” narrowed down to project or research specific data:

> For me data is anything that I am analysing, or using to train a system. Machine translation systems are trained, so in that case the data would be a parallel text, a text that is translated and you have the source and the target language, and you know that this sentence is a translation of this other sentence. Then you have the main parallel. If you have a spreadsheet, for instance, you could have column A with the source text and in each row you have one sentence and column B would be the translations, in each row you have the translation of each sentence in the source text. Instead of using excel you use raw data, you use just a text file.

Others were less expansive in their definition of data, but their response nonetheless indicated the influence of their disciplinary background on their perception of data:

> because I do come from statistics and maths and stuff, to me data is numerical. Numerical or categorical, so I would, in my head, I would mostly view if someone was talking about data as just like a spreadsheet or something.

This was discernible elsewhere in the numerous responses that cited data as “text”:

> But it can also be text.
>
> Data for me is text. So it’s language and it can sometimes be in a raw format or it’s annotated with linguistic information.
>
> Em, yeah if it comes, if it comes down, I might use the word data actually. Um, yeah, well, a corpus, of texts.
>
> So, I would use the word data then. It's either raw text, or a corpus, a corpus of raw text, or linguistically, or linguistically annotated data and that's often xml, I suppose.
>
> The data I work with is text.
>
> Just text itself is data. But someone would have to specifically say, you know if they were going to start using data, they would have to, in that context, they would have to say it first.
>
> I think, a data dictionary would just be like, yeah, if you, I mean you probably don’t wanna know this sorta stuff but if you’ve got, like say, a categorical thing and there’s like three or four different categories in that column, then you want to know what each of those categories matches up with because people, I mean, when I name my variables I try and make them descriptive, as much as possible, but some people just put like a single letter and it’s like meaningless.

And for one researcher, data could be either text or numbers:

> A spreadsheet of numbers, words, it can have, you know, they're either like integers or they're, whatever like, decimal numbers, or just different kind of categories and letters and stuff. But the people then would use data to talk about text. I mean you could say that text is itself a data and that's the data that I work with the most.

It was observed that data can be project dependent, with "the definition of the data I'm using depend[ing] on the experiment I'm running right now":

> I think data, different researchers will define data different ways. For me data, the definition of the data I'm using depends on the experiment I'm running right now. Because depending on what I'm doing I will use different types of data. It's such a broad term.

This in itself can be problematic, particularly when it comes to personally identifiable data or issues surrounding data privacy, as the definition of data can change over the course of a single project wherein "you're literally changing the dataset with every iteration":

> Let's say I'm working on an ontology, which is essentially a vocabulary, and there's various terms for just some very specific meaning. And tomorrow I say Oh wait a minute, this design doesn't make sense, or This is not the right way to approach it, or I can do it in a better way, then you just remove it. So, you can see a vocabulary as a kind of a dataset, so you're literally changing the dataset with every iteration. And someone else using it then has to be able to reference the correct version of the dataset. So, if tomorrow you're using a bunch of privacy policies and someone updates them, then you have to say that OK, this is no longer valid, I'm going to create a new one with the new privacy policy and then, rerun my work over it. So, at some point you have to, like, stop and say that I cannot incorporate new changes anymore because I have deadlines and so on. But, ideally you would just keep on incorporating changes as much as possible. But then that brings in the question that are you aware what kind of changes happen over time?

Many of the respondents cited data as a modular concept, with one specifying that data can be domain specific:

> For us it will be the collection of the corpus that we call as a data and then we choose the domain of the data, is it from coming from Facebook or is it coming from Twitter which area we want to choose and sometimes the collection of data is also different in all the resources available for that domain like Twitter.

In addition to being domain specific, data can also be discipline specific or specific to your profession. The following excerpt displays an interviewee's awareness that what is data for them may not be data for someone with a different profession to them:

> Maybe a doctor is only interested in things like your age, whether you are a woman or a man, whether in your family there are some cases of cancer. So, more statistcal,

> just the numbers, and Yes and No questions, rather than all those things that I am analysing.

One interview in particular was notable for the stress the interviewee placed on what they referred to as emergent or a-typical new modes of data found in humanities research environments or cultural heritage environments. This they contrasted with the more "traditional" data such as those found in the computer sciences:

> traditional, you know, computing science data, which is probably, you know, bits of code or mathematical statistics and that kind of stuff.
>
> for me, my data, I guess, would be multi-modal content. So, like, for example, archives can be, you know, written paper, images, video clips, like there's all these different modes of content. So, I guess, I would consider that data
>
> I think that'll be one of the challenges of working with, you know, cultural heritage materials and saying, like, this is data

One interviewee noted a change in how data is conceptualised, noting that data used to signify something "more general. It's just stuff saved on your computer somewhere or on the cloud" but now it specifies materia that has been structured and that has the potential to deliver insight:

> I think is these like latest few years, last few years, people now refer to data as like more like, not big data, but like data that you could use to analyse and to get some insight out of it. Rather than, I feel like the term data is more general. It's just stuff saved on your computer somewhere or on the cloud. What now when people say data, it's like something that is a bit structured and you could get some insights from it, like if somebody comes to me and say I have data. I would think, this kind of Ok, he has some kind of structured thing, and he could pull some insight.

This was backed up elsewhere in other interviews, with data as structured, sequenced, or as specifically organised material with a specific context being mentioned on three other occasions:

> It's called like semi-structured data.
>
> So yeah from the data I would definitely say it's a series of sequential events
>
> So whatever part of data that you get, you have to be aware that that data that you're using is for that specific, let's say, domain, for that specific genre, for that specific time.

In contrast, two other interviewees expressed the divergent opinion that data does not have to be structured or sequenced, and can in fact be "as good as random":

> data is not, it doesn't have to be like sequential or it doesn't have to be like there's any kind of causality between it. It's very, I think of it as a very unstructured material

> data doesn’t have a meaning to it. It’s definitely like, if it was to say more that it’s meaningless, you know, without context, like this dark data stuff, you know, without context, it’s just, it could just be, it’s as good as random.
>
> I mean data doesn’t have any meaning in and of itself. It’s just sort of is, and then you kind of, you try and only apply as much sort of meaning as is logical.

One of the most interesting themes discernible across all answers to this my request to define data was a tendency towards terminological comorbidity, particularly when speaking about data as a structured entity, with data and information frequently being equated as coterminous and data referred to as “information,” “stored information,” or “any piece of information”:

> Whatever stored information that I can manipulate, search, query, get some statistics about. They can be like sensor recording queries, user activity, document text, image, whatever. I mean whatever you can actually store somewhere and then look into.
>
> I would say, data would be, like, information that could be quantified.
>
> So, data in any case is information, whether you give it explicitly or not.
>
> Basically, data is any piece of information, literally anything, but if you're looking for a computer science point of view, any structured bit of information is data. So, you can have a binary block which in normal speak would mean it's just some file that runs and does something. So, you may not be able to read it, but it doesn't mean that it's not data. Like, it's not identifiable data for humans but machines can obviously understand it, so it's still data. So, it's still structured information. So, anything that's structured information is data.
>
> I mean, I suppose, it’s any you know, any piece of information that can be you know that can be recorded in an index, I think in some way or other on the computer
>
> I have to think about it, because I haven’t thinked like it, we call data rather like a corpus for us, it’s a piece of informations and where data and we did the annotation.

The potential for confusion is not confined to the instances of terminological comorbidity outlined above, but rather to the very concept of “data” in and of itself, with one interviewee pointing out that just because someone else identifies their data as “machine processable” does not mean that their “machine processable data” is the same as someone else’s “machine processable data”:

> Or, I'm just saying Oh machine processable data, as people write in their papers. But then something comes along later that says OK, this is also machine processable.

The same can be said for personally identifiable information, with that which is considered “personally identifiable” being subjective, open to debate, and liable to change:

> So, someone else's definition of what can be considered as personally identifiable may not be as extreme as someone else, in which case they have a problem if they're working on the same set of data. So, if both of them are saying OK we are going to remove all identifiable pieces of information from this data, one person's dataset is going to be much larger than the other person's. So, sometimes when, like in that case, you're literally changing the definition of what is data.

This points to a systemic problem relating to the overdetermination and oversaturation of terms, with one interviewee pointing out that there are an abundance of everyday but nonetheless important terms that mean different things to different people:

> it would be exactly the same word, and it just means different things. Like classification for example. It's a very general term it's like What's classification? Well for example, in my group here in Trinity classification is usually, when you say classification people think Machine Learning classification, categorising things into multiple classes based on a Machine Learning model. Which is a very specific use of this word but like in other places classification is just like, I don't know classification of clothes or products or anything.

The status of some "data" as data is assumed implicitly, but not everyone identifies their data as data, which has worrying implications in relation to consent: if someone does not realise they have given up some facet of their data, how can they consent to it?

> So, actually you have various definitions of data that people implicitly assume, but they just don't term it as data.

That this vagueness is intentional and deliberately encouraged or cultivated, at least for some, was outlined by one senior computer scientist who commented on the resultance of some companies to name certain materials because to do so would make users aware that some facet of their data was considered an "asset" and therefore was of value:

> it's often very revealing how different people view the same things and what mismatches go on. I mean it'll come up in in any sort of systems analysis, you know it's a quite a common issue but what's interesting is that we're ultimately trying to map it into something that we do need to give a name, you know. So sometimes we being, you know you have a debate, oh you can't call it that because, you know, this other stakeholder will now see it because it's explicitly labelled and everyone here, they don't want to be told or the people using it don't want the users to know they treat their comment their post as an asset, you know. There is often sort of sensitivities in there. So you know, yes, it happens an awful lot and that's sort of part of the you know, it's part of the process and actually, you know, what's left to decide who is the most important, who are we really primarily targeting the bit of work at, and then sometimes you may take that and say well now we'll recast it a little bit and then show it to another person in the value chain so, and that's sort of like a lot of my work is typically interested in problems that have a value chain in. So there's a customer, a provider and then a supplier at the very least and sometimes it's a lot

> deeper than that. So you've got like you know different levels of interest of what the problem is.

Lastly, before moving on to examine the reactions to the presented a timely reminder that data itself is subjective, and even the most diligent of researchers will have a subjective opinion on what data are:

> At the end of the day we are always looking at this from our own perspective, from our own experience. So, I will always analyse data thinking of how I use data in my research. And I won't understand what other people think is for them data.

Reactions to presented data definitions.

Responses to the data definitions presented to interviewee participants (See 7. Annex 1 Section 2. Question 5.) was polarised. Three interviewees made the insightful observation that these definitions appeared to be philosophical, anthropomorphised or even poetic:

> It feels a bit like, kind of like philosophical definitions, but a lot of it is true. Yeah I'd like to think about the practical side of it, I mean it's really nicely written. It's kind of, yeah, it's a bit more on the philosophy level I feel, from kind of my point of view.

> Could be, I guess, you know, if you're a philosopher probably that could be a PhD thesis on its own, how to define data right?

> it's almost like, what do you call it, anthropomorphising data or something. Like, you know, "data has no truth." I mean, it's very philosophical. I mean I guess I'm just, I'm okay with, you know, intuitive kind of descriptions of things, but I wouldn't…these are not descriptions that I would, I mean you could say that anything is a fiction or an illusion. I mean, I would say by comparison that data exists, it does exist, it just exists in of itself. But that's kind of far as I would go with sort of abstract metaphorically kind of descriptions of data.

> I would imagine like there is kind of like diversion in how the people define things. It happens a lot in computer science. A lot of people have like multiple opinions about different things and what they mean. But yeah, I would…like this is a bit… Yeah, this is something I don't really expect usually from somebody who is working in computer science. It could be somebody who is like a poet and a computer scientist at the same time, but I would expect usually a kind of difference in the way… Because like I think it depends on how people have worked with a specific term before, or how they use it, and even in different workplaces for example, we use terms totally in different meanings.

This last excerpt contains one key observation that, to my mind, marks out the distinction between data as it was largely defined throughout the material covered in the literature review, and highlights the importance of these interviews in terms of bringing together a richer understanding of how data is defined by practice based researchers in the field of

computer science and computational linguistics: how one defines something depends on the workplace and whether or not they deal with this item or concept on a daily basis: “I think it depends on how people have worked with a specific term before, or how they use it, and even in different workplaces for example, we use terms totally in different meanings.”

This problem of overdetermined terminology, or terminology having multiple and variable meanings was presented by two computer scientists as symptomatic of their discipline as a whole:

> it would be exactly the same word, and it just means different things. Like classification for example. It’s a very general term it's like What's classification? Well for example, in my group here in —— classification is usually, when you say classification people think Machine Learning classification, categorising things into multiple classes based on a Machine Learning model. Which is a very specific use of this word but like in other places classification is just like, I don’t know classification of clothes or products or anything.
>
> I: Can you give me an example, just because I don't...?
> R: Like people talk a lot about “agile.”
> I: Agile? OK.
> R: Yeah, just agile, and agile development. You have to be agile. This kind of, it’s a bit of a vague term, especially to me, and sometimes this is like this is a bit more like when people start to kind of marry technology and philosophy, you get some kind of weird things. So, like for me I understand that OK you have to be agile as in you have to be able to change quickly and rapidly and adapt to the new environments and all that, but like people I feel like people have different opinions and they would define this in a thousand ways and like big data, for example, is a big thing like I don't people what do they mean by big data how big it is. Is not really defined, some people might think that 10,000 documents is big data, some other people say No actually that’s not big enough.

It was observed by a number of interviewees that consensus regarding data could perhaps be reached on a modular level, among smaller research groups or “within a contained group.” This sentiment was echoed throughout several interviews, with the following two interviewees commenting on the necessity for scholars to be “clear about which definition you're using, or how you're using it” and perhaps even be prepared to alter that definition depending on who you are dealing with:

> I think within a contained group you could solve this kind of misinterpretation, or like what people mean with something by just explaining things and just making it clear to everyone. I think it's more about what people understand rather than the term. If everyone understands that, I don’t know, paper is actually this thing [holds up a pen], then it is fine, as long as they all like agree that this is it. So I think, like from my experience it causes problems but like after a while people learn what do people mean by this.
>
> I think academics love debating all types of terms, even just, you know, even in one discipline they all debate about one single definition of one thing, so, I think data is

> probably a huge thing if you're using it across the disciplines of, like, no one will ever agree on one definition, but you might get a group of scholars who agree with one version of the definition.
>
> I think it would be up to the scholar, for example, if you're in the Digital Humanities space, to figure out which definition to use for which group when you're speaking to them. But also, I think, when you write anything, if you're just clear about which definition you're using, or how you're using it, I feel like it's hard for someone to refute your whole work based on a definition as long as you're clear, like, this is my interpretation or which, how I'm applying it. I think as long as it's clearly laid out then it's easier.

A particularly self-aware and reflective interviewee, with clear-cut and well established research methodologies and a commitment to researcher due diligence and best practice noted that assumptions regarding what data are are inevitable, because practice based researchers will always approach the term thinking about what data are *for them*, what constitutes data *for them*. The inevitability of "looking at this from our own perspective, from our own experience" can be countered by clearly outlining your assumptions on data at the onset of a publication or research statement, or when communicating with others about your research and your data:

> Because, yeah it will all, at the end of the day we are always looking at this from our own perspective, from our own experience. So, I will always analyse data thinking of how I use data in my research. And I won't understand what other people think is for them data.

The same interviewee similarly remarked that data can mean different things to different people, observing that the data they were interested in may not be the same data a medical doctor is interested in, but that both are covered by the term "data":

> Maybe a doctor is only interested in things like your age, whether you are a woman or a man, whether in your family there are some cases of cancer. So, more statistical, just the numbers, and Yes and No questions, rather than all those things that I am analysing. So it could be problematic, you have a challenge there, define data...!

One interviewee remarked that the definitions presented a "very negative connotation of data" but that this caused them to reflect upon "the limitation of data itself" so as to become more objective in their research:

> it seems to have a very negative connotation of data, so yeah it's something that make me reflect actually, because some of them are true I mean when you manipulate data you should always consider the limitation of data itself because then you can draw conclusion that are a bit more objective. Otherwise it's like you can direct it towards false assumptions and forget, you know, about the bigger picture because it's true, data are just, it's like a snapshot of some phenomenon.

In terms of the presented definitions that met with resistance or rejection by the interviewees, one strongly disagreed that data “cannot be questioned”:

> So, ummm, data cannot be questioned. Ummm, I don’t agree with that. I think everything can be questioned, because after all, whatever data that you get is just like a little bit of what is happening. If you think that we work with natural language, like in my case, I work with natural language processing and language is moving and it’s, like, it’s always transforming and there is always things, like, words that are not used anymore, or new words, or slangs, there is always a process, it’s like, language is not like a block that is stuck, not like a wall, you know, it’s something moving, something alive.

There were similar, reactions to the statement that data has no truth:

> It has, like, like my example, if I give you data and I tell you it’s from one system and it’s from the other one, the data still has truth, not the truth that I want it to have, but it has truth. So, it really, there is always the truth in something, but whatever you claim with it, that could be not truth.

> So whatever part of data that you get, you have to be aware that that data that you’re using is for that specific, let’s say, domain, for that specific genre, for that specific time. So, it can be questioned because if I use the data that, like if I take literary tests from the, like, 10 years ago, I can, and I claim, look how people are using the language and I can claim, no actually, that’s how people were using the language back then, but in 10 years it could have changed, so we have to do something else now. So, I think data can be questioned, but all the data has its truth. Even then if the text that I take from 10 years ago, it’s not the truth for now, but it has its truth for 10 years back in time.

> I disagree with this. That data has no truth.

> Because for me, data is like the truth, all the truth are there but it is our way to modify or to summarise those data that makes them looks bad or good or fit or not fit to our research.

> Data, all data, has, they are truth. Even if it’s false or not, for example.

> I don’t think data itself are false, I think that this one is really true: always pre-constituted and shaped by the parameters for their selection.
> I: This would of course make sense given that you are working with search results.
> R: Definitely, because it’s like data are false? Data are data!

> Data has no truth? I don’t think that’s true. Data, all data, has, they are truth. Even if it’s false or not, for example.

One researcher argued their point by demonstrating that even data that is presented as “false” is not itself really false: it is still data, but the person responsible for it has behaved unethically and made false claims about the data:

> I'm surprised by the false, data is false. I don't, I think I need more context in that one. Well, data could be false if you make it false, you know, like if I tell you that I need to use, I don't know, translation from an NMT system and I give you, or I take data from, like, another, a completely different system and I claim that to be what it's not, that could be false. But it's not the data that is false, it's what you claim about the data that is false.

This opinion was echoed in the response of another interviewee who again noted that, in the case of "falsified data", it was the agents, and not the data itself, that was responsible for the modification or falsification:

> I believe that data is something even so if we like change it or use it for our purpose to make someone believe in our, so it is like modifying or falsification or fabricate the data. It is like no, no data is not that because data is something given from the nature or from somebody else.

In contrast, others agreed with the statement that data has no truth, commenting that what data are and the value or meanings attributed to data are done by others; there is no inherent meaning or value to the data in and of itself:

> I completely agree with this part, it has no truth.

> You know that the data is everything that you saw. Then every day, how can I say, I bring some, for example, data, some sample for showing that, for example blue is data, pen is data. But it doesn't have any meaning. Information, meaning the, means that, the meaning of the data. Where your data has meaning, it means that now they are valuable for you. Okay?

> No it does not surprise me and in that case, maybe I need to read more but so far the few words which actually I saw here it's definitely goes with mine, because it's pre-factual, it's something we call it, it can be like social stance, it's something. When you, it's how we, we have social bias, social stance, and depending on what is my social stance and depending on what I say, so if you don't know that fact before, you cannot understand me. But that's where, so in this way, these words are not like, rhetorical, it's very common. It's hyperbole, when you use a hyperbole, it's very common in sarcasm like you say like innit, you use the word.

For a number of the researchers, what data are was based around perceptions; our perceptions of the data may change, but the data remains the same:

> No, not all around this, I think it is all completely spot on because I you know that whole question of you know, can you extract truth from data. Like, data is just evidence. I mean the truth is completely based on who you are talking to, you know, what the question being asked is and then you know, these will propagate through, you know, exactly what, you know have you selected and selected that data. So that's, I think, absolutely the you know crucial part of what you're, you know, what

> your sort of data analytics and knowledge engineering is being able to really show what the providence of our processing of that data, so what did we select.
>
> It's like our interpretation of data can change a lot, and it change in terms of how data has been collected, and what kind of conclusion I get from that. So it's like more the conclusion itself can be false or right and it's strongly depended from the parameter that I used for selecting that and…

There was one particularly negative, if not hostile, reaction to the presented definitions:

> all of these are just nuts like, "pre-analytical, pre-factual"…euk, what? And then, "has no truth" is pretty strong, yeah "neither truth nor reality," "fiction," "illusion." I mean, "fiction illusion"…I understand that models,models are fictions or illusions for sure, but data doesn't have a meaning to it. It's definitely like, if it was to say more that it's meaningless, you know, without context, like this dark data stuff, you know, without context, it's just, it could just be, it's as good as random. So perhaps I'm just not as magical with my descriptions.

Finally then, one researcher argued that many of the presented definitions of data were, in their opinion, reflective of a "lack of understanding of data," of its purpose, and its provenance:

> I think that some comments can come from a lack of the understanding of data.
> I: Yes.
> R: And the purposes of the data. And the research. And the source. And the curation of it. And all that sort of stuff.

The onus then lies with the researcher to conduct their research responsibly, to remain clear about "about the data, the sources of the data," and "the test data":

> It's a whole lack of understanding about data, how it's created, the source of it, and so then you have, you'll find, you could have scientists but they're not really, or engineers who are creating a system but they don't even understand the importance of the selection of the data. Alright? So sometimes you need to be biased in your collection if it's for a specific purpose. But, if you're claiming that something is broad domain and is able to cope with language in general, then your source should be covering all this cross section of language. And so there's a lot like in our field if you write a scientific paper, you've to be very clear about the data, the sources of the data, and then the test data. So if you're reporting on the quality of your system you have to be very clear about what was the input and what's your test data.

4.2.3. Data cleaning, pre-processing, and processing.

Data pre-processing and processing

In the majority of cases throughout these interviews, the terms data pre-processing and data cleaning were coterminous. The coterminous nature of these processes was even made explicit to me in one particular interview:

> some data comes from Twitter and has a lot of strange characters, or emojis, or like people sometimes write, like they would write one word and then repeat the same letter like I don't know 10 times as like emphasising or something.
>
> I: Like in speech?
> R: Yeah, exactly, so trying to get those and basically fix them and return it to the…
> I: And is that like preprocessing or cleaning or?
> R: Yes it's cleaning and preprocessing of the data.

Interviewees with backgrounds in computational linguistics outlined established, discipline specific procedures that you "usually perform" for the pre-processing of data, such as the "natural language processing pipeline":

> Most of the time, I work a lot with textual data, so it means that most of the time I do some, what is called preprocessing of data in terms of, there is a work natural language processing pipeline that you usually perform on text.

This pre-processing, that is the type of pre-processing done, can be both problem-specific and researcher specific:

> for example, working with text, textual data there is some pre-processing that we would do, depending also on the problem.
>
> Em, there are many and it's just, you know, your personal preference but guided by experience that make you choose some specific. More than programs, it's like which kind of technique you are going to use, because you want to recognise what kind of language has been used you want to split terms like a word string in the single words that composed the text. You want to do some kind of analysis in terms of removing noise like misspelled words, you want to remove maybe some kind of characters that are maybe not informative like punctuation, or some words that are so common that they don't bear any kind of information.

The type(s) of pre-processing adopted are also specific to each language, and it is up to the researcher to recognise which one is required:

> these are all language based, many of them are language based so you need some specific technique for each language, and then you apply some tools for example indexing these words, so you can search on them very easily, see how many words that are there in the dataset.

In terms of workflow for the pre-processing of textual data, the following two excerpts outline two different, but equally methodical and clear, approaches towards the processing of data:

> So, first of all we need is parallel corpus. So we need the source language and the target language, whatever language is to be aligned in a sentence level. So, generally the data, like if you try to crawl like texts from the web, they are never aligned. So alignment is one of the things that we have to do, and cleaning the text, like getting rid of, normalisation, getting rid of accents and punctuation and all this kind of stuff. But I think the alignment is one of the most difficult things that we have to do, because if the alignment is not very well done, the translation is not good.
>
> But I generally organise my data as source, MT, post-editing, and reference, if I have a reference, for example, the reference would be if I have this text and I ask for a human translator to translate it, that would be my reference. And then I have, I use this in a machine translation system, and then I have the machine translation output, and then I ask a translator to look at the machine translation output and fix it, and then I have the PE, so I can organise like that. Sometimes, we don't have the human translated part, because there is no gold standard for the translation, so what I do is organise source, machine translation, PE, and my PE is also my gold standard because some human looked at it. And then I can have the source pre-processed and post-processed, or not and the PE, I can have post-edited the clean part, or also with the annotation for errors or not.

Interviewees identified the pre-processing phase as a phase wherein the data gets classified, but also where they are free to come up with new classifications depending on what they see in the data:

> So, normally you start out with your background knowledge and your preconceived ideas of categories of content and that sort of thing. But normally through a pilot phase or a pilot test, you'll discover categories you didn't know exist for example, and then you add them to it. And then you have a pretty good matrix structure that you can work with, and then from there, like 99% of the data will be able to be categorised and catalogued and then you might find some uncertain data that you're, like, I'm not sure so this go in the other category.

Elsewhere, among the computer scientists interviewed, two observed that, in the case of data that is text-based, and in particular of difficult to process pdf files, the pre-processing phase sees this material manipulates so as to make it more accessible and searchable; it's the same data "in essence," they argue, just made "usable":

> There's different kinds of material or data that I use. So, some of them could be something I directly take from a website. Literally just copy text, a bunch of text. For example, I was looking at the actual legal text of the GDPR and it is online, which is good, because these days, it's like everything is online. People can say Oh here is the url for something and you can go and access it. But then it's in a form that you cannot refer to someone, I cannot tell you Oh go look up Article 18.2 whatever and you, being a human, know that you have to visit that url, scroll down or find Article 18.2 and see whatever is written there, but how, for a machine this is not trivial. You can't just say Oh do this the same as a human would do. So, then I had to look at how do I transform that legal piece of text into some way that machines can understand. So, then I had to break the individual articles, points, sub-points, give

> them each an identifier. So, in essence, it is the same data but now I have added additional annotations which will allow you to point towards different bits of article. So sometimes you get data which is ok and fine and what is the standard way of representing it, but you can't use it in your research, so you have to annotate it additionally in order to make it usable.
>
> We look at the official sort of HTML document from the EU, which is you know, the actual reference for somebody making a judgement on the law and they will have it, usually they'll have a text version, a PDF version and a HTML version, and that's then I'll canonical version underneath generating all of those, and you look at the HTML version and it's sort of done as a table okay, so it's got all the sort of clause numbers and then the clause there which is logically from a parsing point of view, you know, it implicitly has the information there but a table isn't the right construct for representing, you know, a set of clauses and a set of some bullets and things like that. And all those materials are available in the HTML, you know, they just don't haven't them properly. So we can figure out a parser that will process that and extract the information but it's, you know, it involves us having to get rid of, we don't need to maintain the fact that this was presented on the web as a table because that was already visible to the user anyway. It's just, I mean it's used an awful lot because it you know does sort of work as a, from a presentation point of view but it sort of destroys a lot of the logic of the document, it just presents the logic but it doesn't maintain it in the actual encoding. You know, so that's, we essentially sort of probably reformat and get back to something that was in the original format. And you can actually, the actual legal text is actually kept in the sort of XML format but I think they publish it.

The importance of accountability regarding annotation and accountability was highlighted across disciplines, with both computer scientists and computational linguists stressing the need alternately for making "an explicit way of referring back to the same bit of text, whether it might be a legal clause or a particular word even in an unambiguous way" or for establishing annotation guidelines or an inter-annotator agreement:

> so we will do things like date the document. Perhaps we may even keep their IDs in it just for backward compatibility but we add our ones in. So we have an explicit way of referring back to the same bit of text, whether it might be a legal clause or a particular word even in an unambiguous way, you know, and it's actually still surprisingly difficult to get exactly right. Especially, if you're trying to do it in a way that will last, you know, over a good period of time because even, you know a lot of people do this but you know, you know the classic joke in our area is well somebody published this data and then it disappears it's cos the guy finished his PhD, you know, he's not maintaining it up on, or on a site anymore and things disappear very quickly. So that's why we like to take copies and then have a clean version of the copy and part of the reason also of having a clean version is it may in future encourage other people to refer to our clean version as the one that they annotate against.

> annotating the data, but before annotating the data I had to come up with annotation guidelines, and before doing that I had to actually do an analysis of Irish that hadn't been done really before in this...computation wise.
>
> An inter-annotator agreement should take place on any type of annotation of corpus, but I find it doesn't, and it's a... It depends on the people, on their background. You'll find if they're computer scientists and they're an engineer, they don't really know or care

While some among the researchers (particularly those working within the discipline of computational linguistics) annotated everything, others only annotated material that was specific to the purpose or aims of their research:

> I: What are the reasons that something would be annotated and something would not be annotated?
>
> R: Oh that's a, it depends on the guidelines you're using.

One computational linguist was particularly eloquent and impassioned regarding the need for inter-annotator agreements and co-operation across disciplines, citing in particular a distrust for the methodologies of computer scientists or engineers engineers who take on a project that involved linguistics, without fully according by or respecting the protocols established by the computational linguistics research community:

> That people might release a dataset that's not reliable or replicable or their experiments can't be repeated not through them being chancers, but just because they're not aware of, say for example some people don't even know about inter-annotator agreement. So that means their training was bad, or they didn't do enough research, or they might be what I would call an imposter in the field. So they're a computer scientist who's like Oh I'll do this Machine Learning and I love this or I've done this before. Oh lets look at this new dataset. And they're not really in that field. So I could draw an example say here someone who works on tweets, great. And then somebody in the Digital Humanities asks them Can you help me with old texts and analysing old texts? And they've no familiarity with that area at all. But, and they just go hell for leather into it without actually getting to know the domain and the text and how the data is presented or whatever. So I think, but I don't think it happens too much but I think there is definitely a small group of people would could be up to that. And not intentionally. Just lack of knowledge.

Data cleaning

What and why material is removed.

What is removed during the data cleaning process varies, but of the thirteen researchers interviewed, the following represent the most common reasons.

i) What is removed is project or research dependent, with one interviewee noting that “It depends on the research, completely. It depends on the purpose of the research.”

Similarly, another cited “data that we don’t need to model” as an example of what gets removed. Data that is immaterial to the purposes of the research:

> We are trying to, maybe the data that we don’t need to model. We’ll remove those kind of things. For example, in our scenario maybe they say that The patient goes by walk to the hospital. It’s not important for me he is going by walk or using the bus for going the hospital. Just for me is the, Who is the user? and Where is he now?, for example. Again, the data that I need to model I will keep it, the other one will be removed.

Again, one interviewee notes that data that “are not important for my work and for my research” will be pushed or thrown out:

> As you mentioned, part of the data will be missed, but I can say that those are not important for my work and for my research and because of that maybe we, push it out actually. Rather than those data, other things will be in our model, just a part of the text that’s related to data that those are not useful for our research purpose will be threw out. Other things will be modelled in our system.

Material that may have been gathered accidentally, through a keyword search for example, that turns out not to be relevant to the project or needs of the researcher is also discarded:

> The one that’s not relevant with my study, for example, it has mislabelled like, the title representing this risk prediction but the content like predicting the transmission of disease, it is quite different. So I will leave out this research and then like some research are related to the stock markets but they use the health domain to those many algorithm of the stock market. So I just leave out. So the, I think the research that is not relevant to my research but it has similar label…
> I: Okay.
> R: …in their key words or their titles.

A comparable example comes in the form of material gathered that, upon closer inspection, proves to be from an unreliable or untrustworthy journal: “Maybe the article that is published in not well eligible journals.”

Certain of the data that can be removed is removed because it is outside the parameters of the research question. For example, the following researcher designed a twitter bot that prompted twitter users to reply with “#yes or #no or Like.” Any response that was external to those parameters was excluded from the study, even if it was clear that the response was a synonym for a “#yes” reply (because the researcher was looking to identify sarcasm, and certain responses were clearly sarcastic):

> Even I got a response like Is that you Sheldon? So people were also being sarcastic in their reply. So, we could not process those so we had to skip those by people they can’t respond #yes or #no or Like.

While many of the interviewees spoke lucidly and understandably about what and why certain material was removed, others were more esoteric. The phrase is an example of the opaque language that can surround the cleaning of data:

> It is project-dependent but ideally, in terms of pure statistics, you would expect everyone to, kind of, leave the outliers out.

When prompted as to what exactly outliers are, the interviewee elaborated that outliers were "would be anything that doesn't actually conform to the model" but conceded that they themselves "have a very weird way of defining their processes. They're very vague":

> Like, I don't have a traditional mechanism of research where I deal with data, and then quantify it into some numerical value, and say Oh this doesn't fit in so I'm just going to leave it outside. I deal much more with modelling. So, outliers in this case would be anything that doesn't actually conform to the model, like they have a very weird way of defining their processes. They're very vague. Then I'd say that OK this is not trivial, I'm not going to be dealing with that.

For one interviewee, material may not be removed, but it may as well have been, because it is "de-prioritized" and as a result "the knowledge engineer will rate those CI but not include it into the calculation":

> Maybe I don't remove it but I put less priority to them. So the knowledge engineer will know that all, there is all this kind of data, this is not and the percentage of this appearance is not so high, so this, yes but I will not remove it either.
>
> So the knowledge engineer will rate those CI but not include it into the calculation.

ii) Errors featured high on list of materials that are removed during the data cleaning process.

> I: When you're annotating them, what type of things would be left out or, what would be removed?
> R: If there was mistakes or errors.

iii) Data is cleaned so as to standardise it. Many of the interviewees clean their data in order to standardise it. This is largely completed using specific software, with one interviewee citing the steps taken in Parlai as "slightly unnecessary":

> the data cleaning that takes place in Parlai like, there's probably like three steps to it or something crazy. It's slightly unnecessary but if the end goal is for everything to be the same, that's a worthwhile enough call.

iv) Others referred to what they do as a "restructuring" process, that may involve changing the metadata, but that they ultimately try to leave the data alone:

> there is a certain amount of restructuring the data, but in terms of changing the underlying stuff itself, it mean maybe you've changed it because the context is different, or it's got different sort of metadata for sure, you might just drop some metadata, some like bits about the data that you don't really need, just so it's more clean, but in terms of the actual stuff itself, you try and leave that as much alone.

v) HTML tags, links, and formatting tags were cited by several as examples of material that is problematic and can often be removed. These items can be considered related to the preceding point, as one interviewee argued that without "strip[ping] out" these original source html tags, the link itself may change and the source data may be lost.

> So the stuff that we take out, so there's, we get a lot of material in our sort of especially a published HTML document that are perhaps tags to do with formatting, you know, which people do really, I mean again, because we, I've sort of come through the process of you know how these things should be published but you know the HTML, the format is extremely forgiving. So, you can do all sorts of horrible things and the browser will still present it in a fairly decent way.
>
> So what we would have, just the core text and just a core set of tags inside it. So we would use the standard HTML tags and then other, any other ID tags we try and control which ones sort of stay in there and which ones we strip out. So that's partly just to make it a little bit easier to manage the annotation of it but also we need to, you know, the reason we do that as well is because we can't rely on the source data staying the same. So if we just point to a, to a you know a really big long document, you know like a legal text on a website, depending who it is but we, even if it's all very reliable like the EU, they may still do an update and you know, do a lot of changes. You know, what you would read on the screen might be the same but there might be a few like highlighted errata, but the actual underlying HTML could be a rejigged quite a lot, you know, so for example often in HTML you will have essentially bookmarks, you know, you like tag a bit of text and you will say this has an ID and the ID typically comes from a database or something. It's not you know, it's automatically generated so, that's something that you can refer to. Say I can, I want to refer to that particular article or that particular clause but if the document gets, if you like, re-indexed automatically and it's only being done for publications, so nobody else is annotating it, then they don't really care that the numbers are different. They just care that they're unique, okay, but if we've separately annotated and put into a number that they've generated that, that throws us off.
>
> The links are not the same so it's mostly a link. So if you go to the link, extract the meme so that is an image, so then you have to do image processing. If I do the image processing, of course it's gonna be needed, but if I don't do image processing I have to skip those and what we find out that, if you just keep the link sometimes people say Yeah it can be the same link, if you are posting the same meme, but it's not, because big link change constantly. So that's why it was not used for us and... That's why we discarded that.

Again, and in line with point i) that what is removed is project or research dependent, the removal of links can also occur because they are not necessary to the project. In the following example, the material the researcher was searching for was text based, and so all links were discarded:

> cleaning wise we only cleaned only one thing which is, we didn't want to use any sort of link which goes to because we only based on text, what is the mood, and what is the context. That is all we want to use So whenever it comes to a link we just discard it.

Given the acknowledged messy or "forgiving" status of the html format, with one interviewee above noting that "you can do all sorts of horrible things and the browser will still present it in a fairly decent way," unexpected characters relating to html or xml tags can impede or stall the preprocessing process, and for this reason may be removed:

> But sometimes like you would run your preprocessing to I don't know remove some characters from there and then you, like you would find like, they, the preprocessing would fail for some reason. There's like, there could be a character that's not included, like I didn't take into account. Um, sometimes for example I assume that file if let's say I'm working with an XML file, or HTML file which is like the web content, and I would assume that they would have like specific tags. Sometimes the tag is not there and you would need to do something about it, or like sometimes it's there sometimes it's not and then you try to kind of figure out how to, you take another way to get the data.

vi) Text can be altered or "standardised" by being made all lowercase and through the removal of multiple letters; the latter being phenomenon that appears particularly relevant to Twitter data where one regularly sees the use of unstable, non-permanent links, and the use of multiple letters. This process is called "Corpus normalisation" ("what I would do to normalise all the corpus I was using beforehand"). The following are two instances where separate researchers cited multiple letters in Twitter data as an item that gets cleaned or reprocessed:

> Now in terms of that there was also another point which people are mainly mostly do reduce the complexity of the model, they compact every dataset into lower form like it's lowercase. They convert every text into lowercase, which we didn't do because capitalisation is also a feature for us. It's like I'm watching if I see some of these and that contrast is also present. So, filtering wise we didn't do too much, we just kept it only few filtering, only two to three filtering we have used. One is link, second one we, if we have multiple like Looo, multiple o and l, so we reduced that so that if we had three of them consecutive we make it in two because in English we didn't find any words which has three consecutive letters the same. So we converted it into two and that's how we shorten our space and other than that we didn't do any shortening or filtering.

> and the second one we discarded we just find out there is one feature also present that lol, when you have this kind of long, so we shorten it because, for us, we did not

> want to emphasis if you having a big lol has, and a small lol, there's too much difference.

This altering of received data sees some crossover between the acts of data cleaning and data pre-processing:

> working with text, textual data there is some pre-processing that we would do, depending also on the problem. It's not always the same, but sometimes we might decide to remove the punctuation for example or just lowercase everything, or like remove the characters that are weird and like for example some data comes from Twitter and has a lot of strange characters, or emojis, or like people sometimes write, like they would write one word and then repeat the same letter like I don't know 10 times as like emphasising or something.

As noted in point iv), this material can cause the preprocessing to fail. Alternately, one researcher noted that this "abberational" or atypical material was going to be "washed away by the algorithm anyway, they gonna be ignored, or like taken as noise." Such material is considered difficult for a machine to process, and can therefore be discarded, or indeed one can attempt to avoid this type of data altogether:

> Well I guess what you can do is once you are analysing the results then you have to mark this as something that was hard for the machine, maybe discard it from the test set. Or if you are choosing, if you are selecting the test set, trying to avoid this kind of data.

One interviewee specified that in order for the "computer [...] to kind of learn," as little as possible data should be removed or altered during the cleaning process: "in terms of data cleaning, people try not to, they're trying to just make, they're trying to leave as much to the computer as possible to kind of learn, I think." Unfortunately they also displayed a lack of consistency here as they stated that in their own research and data cleaning scripts (which were built by someone else), "there is a list of commonly misspelt words that are kind of like fixed and replaced":

> I'm using someone else's data cleaning scripts but there is like a, there is a list of commonly misspelt words that are kind of like fixed and replaced, but generally speaking people don't do that.

vii) Misspellings are often, but not always, corrected or identified as "noise" and removed:

> You want to do some kind of analysis in terms of removing noise like misspelled words, you want to remove maybe some kind of characters that are maybe not informative like punctuation, or some words that are so common that they don't bear any kind of information.

viii) Punctuation, while acknowledged as useful, can cause problems, and for this reason, as the following researcher indicates, it can be removed, as can accents specific to certain languages:

> it's not because it's not useful, it's because sometimes the machine translation system doesn't deal well with punctuations, so what we can do sometimes, I don't know how is the neural, because neural is a new type of machine translation that is coming up and I didn't get hold of how it works properly yet, but I know in some machine translation systems we get a read of the punctuation, for example, English has no punctuation right? Like, I mean like accents, that's what I mean. And, like in Portuguese we have, like, a thousand of them, so we decide to get rid of them and then it would do up a post-processing, then put them back. Yeah, but we put them back, but for the training of the system we have sometimes to remove them because, you know, like, the machine doesn't get, or sometimes the accents are different and, like, if you work with user generated contents, sometimes people don't use the accents in the, in Portuguese, you know, because, like, you're tweeting, you don't want to, you don't care if there is like an acute accent there or not, so then when we receive without, our machine can deal without as well. And then we just put it back, if it's the case that we need.

As outlined above, once the material has been processed, these "removed" entities will be put back and returned.

Other language specific idiosyncrasies regarding the punctuation of text or numbers can result in inaccurate translations; a point that may be known to a computational linguist, but not perhaps to a computer scientist or engineer with little or no background in natural language processing:

> Or even stupid things like numbers. It's not the same how do you put the thousands and the decimals in English and in Spanish or French. And that's something that many many people—it's surprising—don't know. So they, they will correct things, even if it was done right by the Machine Translation system because in their brain it's hard-coded differently.

Similarly, accents will cause problems across different systems, and so "normalising" a corpus helps to maximise information retrieval:

> The first time that I worked on this topic, information retrieval, I was just adopting the standard pipeline of preprocessing of data, and then I noticed that my system was very low performing and then I recognised that, for example, there was a bunch of words that were very not informative and they should have been removed. So, for example, I added some more stop words to remove from the process or, in that specific case for example, I noticed that the system wasn't able to answer to with some queries, and I looked into the data and see what's happening there, and then I notice that actually there were many queries with Spanish words, and the coding used for the accent in Spanish words was different from the coding the I was using for representing data. So basically I was never able to find that word in my dataset because it was a representation mismatch. So that's some insight that I got after I looked into the data, looked to the failure of the system, and saw OK this is actually not the problem of the algorithm itself it's just a problem of representation and, you know, in computer science you have these different coding system, and if you

> choose the wrong one it’s like you are not representing the data in the same way that is comparable.

Certain algorithms also function better with materials removed:

> for some cases we know that the algorithm would perform better if you remove prepositions or punctuation something like that, and in other cases it might be the opposite we need to kind of have the exact some text because we would be comparing with a text that’s formatted in the same way. So it's depending on what you are trying to achieve and the algorithm that you are using.
>
> But for like if you are doing some like part of speech tagging or something like that, it's just like they don't have any value there and it would just be it might actually be extra noise and make the algorithm not perform well

The process of tagging and alignment was recognised as a difficult one, particularly in the case of translation project that have acquired material from the internet; such material was considered messy and difficult to align, and structure into normalised data:

> So, generally the data, like if you try to crawl like texts from the web, they are never aligned. So alignment is one of the things that we have to do, and cleaning the text, like getting rid of, normalisation, getting rid of accents and punctuation and all this kind of stuff. But I think the alignment is one of the most difficult things that we have to do, because if the alignment is not very well done, the translation is not good.

ix) Material you cannot agree on or reach a consensus over may be removed or marked out from the surrounding data as inconclusive or uncertain. As the following excerpt makes clear, this is particularly relevant to cultural data:

> Sometimes people can’t agree, so controversial issues don’t get agreed and get marked as things haven’t got a consensus around it. You also see for example that you know it doesn’t try and like Wikipedia doesn’t do much translation okay, they say well they do it in certain like physical sciences where there is an objective agreement on, you know, maths or anatomy or something like that but loads of other things they say well no, different countries and cultures should find their own version of Gallipoli or you know, World War II or something like that because they will have their own, you know, genuine perspectives on it you know and even within a single language, people are going to disagree about it and the process of writing a Wikipedia page is almost a process of deciding what you’re going to leave out because you can’t agree on it.

Similarly, uncertain data may be removed, or, if it cannot be removed, uncertain data may make the material as a whole unusable:

> Like, if you can remove the uncertainty, then it's fine. What happens if you cannot remove it? Do you just get another dataset? Do you ask someone else to do it? Like, sometimes you have an uncertainty, and you cannot use the data but you also cannot remove it, so it just sits somewhere for a while.

As will be further discussed in Section 4.2.4 Messy data/ difficult material, uncertainty in data can influence the reliability of the data. it can also impede the standardisation of data, or the move towards structuring or serializing the data. For this reason, uncertain data may be removed because, as outlined below, “you want to get rid of that uncertainty, you want everything structured proper, everything states something reliably”:

> Uncertainty in data. I don’t think it would ever be good, because that essentially means that you cannot use that data reliably, because it's uncertain. So, as much as possible you move towards structured information. Like, uncertain data is chaotic data, you don't know what's in it. So, you want to get rid of that uncertainty, you want everything structured proper, everything states something reliably. So, ideally you want to get rid of the uncertainty as much as possible.

Similarly, contradictory data may be classified as “outliers” and removed, or it may disqualify itself from inclusion by virtue of the fact that it is contradictory, and because the researcher doesn’t “deal with contradictory data”:

> again, depending on the context, you would either classify them as outliers and remove them as Oh this is contradictory data and we don't deal with contradictory data.

x) Material that cannot be understood by the researcher (as opposed to by the machine) may also be removed. This is particularly relevant in the case where the content can be particularly esoteric, inventive or idiosyncratic, such as on Twitter, and where the researcher is working in with non-native language they are not familiar with:

> So yeah and then, like, when I was working, when we working with the tweets, there are things that we can’t understand, you know like tweets that are, like the source is in English but then, like, you know those tweets that you have no idea, that, what people, because I’m not a native speaker of English I ask a native speaker, they also have no idea what’s going on, so, we just decide to get rid of them, because if we can’t understand, we cannot expect the, a human to translate that, and then the machine to understand what’s going on. So, yes I think every data has its particular, you know, challenges, but there are things that, many times things that we were not expecting at all and they just happen

xi) Lastly, data can be cleaned as a result of efforts being made to anonymise the data. However, this is easier said than done. As one interviewee made clear in the excerpt below, it is getting harder and harder to truly anonymise data:

> You know, so we look at, so we look in specific incidences in that data, that sort of data, that’s also that’s why I was asking about sort of reduction, what’s interesting about linguistic data is that it’s extremely difficult to really de-anonymise. So just removing the person’s name doesn’t work very much. If you’ve got a big enough sample of how somebody has translated something or how, if you’ve recorded their voice or something like that, it’s, if you record that broad data, it’s nearly impossible to completely anonymise it if somebody else has another sample of, you know, that

> they can match against and the technology allows you to do that matching increasingly easily. It's, you can't treat that as anonymised data which means you then have to pose a lot more controls about how you manage that from a personal view, so we're sort of looking at those issues currently and so that's a sort of recent research area, especially with the law changing in Europe with the general updated Data Protection Regulations coming in.

Further still, as was pointed out by one computer scientist with a specialisation in GDPR, our understanding of what may be considered "personal" and what may be considered "identifiable" can vary, and this created huge problems for those attempting to anonymise data:

> So, someone else's definition of what can be considered as personally identifiable may not be as extreme as someone else, in which case they have a problem if they're working on the same set of data. So, if both of them are saying OK we are going to remove all identifiable pieces of information from this data, one person's dataset is going to be much larger than the other person's. So, sometimes when, like in that case, you're literally changing the definition of what is data. Like, this data may be a part of some other dataset, but how I derive this data from the bigger dataset itself defines what the smaller subset is. So that's, like, how do you define this relationship between different pieces of data? That needs to be clear. So, that's where you have gaps in communication. That's why you have legal text that should very clearly and unambiguously mention what it is and so on.

What happens to this excess or removed material?

Researcher's reactions varied in terms of whether or not they acknowledged the removal of data, or stored original versions of the data they modified and worked with. One did not acknowledge the removal of data ("No, no, we don't mention the removing the data.") because the data, they argued, was not real or computational:

> As we, I don't have any real, how can I say this, computational data, I mean that the real is stored data. Just I have some scenarios in my mind, or on the paper. Just they are removed and put out. We don't save in anywhere in the data. We have the scenario somewhere, as a written scenario, but in our model we don't have those data, and maybe not keeping out just storing in the, for example, the written scenario.

Others outlined meticulous procedures regarding transparency of provenance, such as "dat[ing] the document," and keeping original and perhaps even useless IDs "just for backward compatibility" and in order to have "an explicit way of referring back to the same bit of text":

> so we will do things like date the document. Perhaps we may even keep their IDs in it just for backward compatibility but we add our ones in. So we have an explicit way of referring back to the same bit of text, whether it might be a legal clause or a particular word even in an unambiguous way, you know, and it's actually still surprisingly difficult to get exactly right. Especially, if you're trying to do it in a way that will last,

> you know, over a good period of time because even, you know a lot of people do this but you know, you know the classic joke in our area is well somebody published this data and then it disappears it's cos the guy finished his PhD, you know, he's not maintaining it up on, or on a site anymore and things disappear very quickly. So that's why we like to take copies and then have a clean version of the copy and part of the reason also of having a clean version is it may in future encourage other people to refer to our clean version as the one that they annotate against.

The majority of those interviewed kept copies of their original data, though not all went so far as to explicitly outline what data had been removed or changed in their publications. One interviewee referred to what you could "get away with" "sometimes." They also referred to a culture that correspondingly does not query researchers on data transformations or removal because "no one actually asks, or no one actually cares how you incorporated the changes":

> Hmmm, sometimes you can get away with publications by just saying we changed to incorporate some reason, and no one actually asks, or no one actually cares how you incorporated the changes, as long as they understand what it is. Sometimes if it's specifically with a lot of statistical datasets you have to exclusively mention how the change was acquired, what happened to the old data.

Another interviewee cited two approaches to this.The first involves stating that items were removed without specifying what they were, merely noting that they were "contradictory outliers." The second involves contradictory data that cannot be removed and therefore has to be "dealt with" and has to be mentioned in the publication:

> Two use cases spring to my mind. One is that you can just remove it from the data, and get away with mentioning We obtained 50 samples and out of which, we found 40 were usable and 10 were not usable because there were contradictory outliers, whatever the reason. But then you don't have to specify what those 10 were, but you still have to publish them as, because you obtained them. And the other one is that there is a contradiction in the data that you have but you cannot remove it. So, if I have some dataset I have to keep it whole, I cannot, like, say that Oh this is contradictory so I won't deal with it, in which case you have to deal with it and you have to mention the contradiction in the publication.

Among the computational linguists interviewed, many of whom had very clearly defined research methods and very clear ideas about best practice, researcher accountability, and the evaluation of good research, expressed frustration regarding the lack of transparency surrounding the pre-processing of data in an interdisciplinary environment:

> There are parts that are really taken for granted, and you will see that also in Machine Translation papers. It was really tough for me to get into the research area, because there are so many things in the pre-processing step. They just, many people writing their paper, that they do some pre-processing, but they never explain what. And I'm was thinking, But that's the key for me! How did you pre-process the text, right?

For this particular interviewee, accountability regarding the pre-processing of data was associated with best practice, but it was not something they always encountered outside of their field:

> When I was doing my PhD I worked with German into Spanish, Machine Translation. And I always did some pre-processing to the German data, and I always explained in all my research papers how I had changed the German data.

Another interviewee stressed the importance of representing the provenance of your data processing, which will be one of the key recommendations of this Work Package:

> So it's about being transparent in your in the provenance of the data processing which is really important. So actually, you know, when we look at this from operational decision making that the tracking of the provenance, that's what we actually use standards for representing provenance and we find that that's absolutely key because you know it's all the steps that you take

4.2.4. Messy data/ difficult material.

Ambiguous data

The most frequently used example of data that can be ambiguous was text based data. Irrespective of disciplinary background respondents routinely identified text as ambiguous, and as something that can be ambiguous "at several levels":

> When you work with text, text is always ambiguous and it's ambiguous at several levels. It can be ambiguous at the level of words, now you know it's not in this case. In this case is more ambiguous the context, because you don't know what was the cause of the crying. But it can be ambiguous at the level of sentence, I mean the same sentence can be interpreted in different way, because maybe the sentence structure can change in terms of parsing. So and that's a problem for example in IR because then you don't know when you are talking about I don't know New York, if it is New York the state, New York the city. I think there is also another city in the UK.

Ambiguity was cited as a characteristic of natural, human language:

> Actually this ambiguity is a real problem that exists in the natural language. Yes, means that nobody can say they don't have any problem like this ambiguity. Because it's a major of the natural language.

> I think when deal with like any kind of language, natural language processing thing is always a problem. It depends more also what you are trying to get, if you are trying to understand what is going on, it's always a challenge, because the yeah language is very ambiguous.

One interviewee with background in computer science and specific research interest in sarcasm detection outlined the layers of incongruity discernible in Twitter text, and concordantly the areas where ambiguity can occur: “So we are finding incongruity between text level, text context, text mood, text social stance.”

Irrespective of disciplinary background the approaches adopted by the interviewees in the face of ambiguous data was twofold and involved i) disambiguating and ii) the application of context.

Researchers from the two fields that made up majority of researchers interviewed were familiar with discipline specific approaches to ambiguity, with one computer scientist that “there is actually a big field of computer science that try to work on ambiguity and to solve it and it’s called word sense disambiguation” and one computational linguist describing a Machine Translation system that endeavours to disambiguate by laying out all possible meanings of a sentence identified as ambiguous:

> So, there are different types of Machine Translation systems. The rule based system is using a grammar and a dictionary, and it would be like trying to teach the computer and code everything that is in your brain, right?
> I: Okay.
> R: And then there they usually have this ambiguity coded and suddenly and then suddenly you have text in different colour, telling you This is one possibility. Or sometimes they have a bar… [...] So then sometimes what they do is they put a bar and then there is like This thing can be translated like this, like this, or like that.
> I: Okay. So, like they disambiguate?
> R: They disambiguate. But they don't know which one to choose. So they give you all the options!

This approach, disambiguation, is brought about via the application of context; which was the second major approach cited by all interviews apropos dealing with ambiguous data. The following excerpt outlines how context is applied to ambiguity to rule out unlikely scenarios or impossible meanings:

> human language is ambiguous, so forms occur, homonyms occur, and then you, then out, a machine needs to, needs to am, contextualise it to find out what meaning is, yeah what the verb is, or in my case then what the verb is, it's plural, singular, depends if it's the subject like "girl" singular, so it has to contextualise the interpretation and then it, it'll remove every option that is unlikely or impossible.

The provision of context can be further delineated, for example in the following excerpt the interviewee outlines rule based and statistical approaches to the provision of context while again highlighting that the provision of context is itself an act of disambiguation:

> Yeah, contextualisation is a, is either rule based, so you say, if it's a noun, followed by two other nouns, then one of those two other nouns are probably not nouns, because three nouns after one another is very unlikely, at least in, I suppose in English and maybe in Irish as well. But em, it can also be statistical, so that's what Google Maps does, if it sees these three words being, meaning that, then in that

> case, when there's a slight difference, it just takes the likely interpretation from the similar scenario. So, if you see those things and you're not sure about one, look at other instances where that combination occurred and take the most likely one and apply it. So, I mean, that's disambiguation.

Numerous interviewees spoke about how difficult it can be to interpret data when it has been separated from its native context and the context in which it has been collected, the parameters and provenance of the dataset:

> Yeah we are trying to resolve it in an automatic way, and that means that it is not one hundred per cent precise, and I mean even in this case, even this is because machine…but even in this case if I read this and then I read this one, I can not say which one is correct because it's like it depend on the context and I am not in this context. So it's really hard to it's like once you get the data and you remove the context, you don't know anymore what…maybe there are hidden parameters and you don't know them, there are hidden variables that, you know, condition the way data have been collected, and once you don't know anymore what these variable are it's very hard to describe.

Computational Linguists and translators appeared particularly proficient in dealing with ambiguity, and outlined methodical and well-established approaches to dealing with the phenomenon, which again involves disambiguation and context:

> So before in the first experiments with machine translation, they would just give translators random sentences and they were expected the machine would do as good as the translator, but the translators weren't doing good because they didn't have context. So now, we keep the context as much as we can and then we test that against the machine, to see.

> what I have is just a randomisation of sentences from the major corpus, right? They have to be randomised cause you don't want your tool to be primed to whatever. But then I'd have to say Well what's the context of this in the major corpus?
> I: So that would be another approach then?
> R: That'd be how I'd disambiguate sometimes, yeah.

In addition, the importance of inter-annotator agreements to prevent or discourage ambiguity in annotation was paramount, with ambiguity at the level of tagging of coding being introduced in one instance as a result of the researcher's failure to create an inter-annotator agreement:

> But then like, they didn't do inter annotator agreement and so well sure, my interpretation of what a complaint is versus...a comment, you know? Sometimes it's vague.

Another interviewee argued that ambiguity only arises as the result of a lack of context:

> It’s only ambiguous if you’re lacking context I guess. I mean there’s semantic meaning in these sentences, like you know, I can read that and you can say Because she is happy, or whatever. I mean data doesn’t have any meaning in and of itself. It’s just sort of is, and then you kind of, you try and only apply as much sort of meaning as is logical. So sometimes people maybe make assumptions and then they don’t realise they’re making assumptions, but generally you would be aware of ambiguity in any data.

This was apparent elsewhere with one interviewee arguing that difficulties in translation arise specifically because of a lack of context:

> when you're working with translations without a context. That sometimes you have the same sentence meaning different things depending on the context. And then you need to like, if you're thinking Yeah we meet at the back, right? It could be translated, well no in Spanish it's also ambiguous. I cannot come up with a word that means two different things. So if you have a polysemous word in English and you want to translate it into Spanish, depending on the context you will choose meaning or the other. If you don't have the context, then you don't know which one to choose.

In addition, ambiguity can be introduced when something is not complete or when there is a partial or incomplete dataset. Again, however, this can be remedied by the application of context:

> Because it is not presented in full context or full sentence.

> Yes, there isn’t enough context and it is not presented in full sentences.

One researcher spoke about the deliberate provocation of ambiguity around and in relation to terminology and definitions of data, and spoke of the rise of “professional ambiguity” in privacy policies and the need for legal text to be clear and unambiguous instead of facilitating the exploitation of personally identifiable data:

> It is very ironic in that you would consider legal texts to be unambiguous, and it is in terms of the intended meaning, but then how it should be used or how it should be expressed in computer science is very ambiguous because you have a bunch of different ways of expressing it, and you don't know which one is correct or which one will actually give you the intended result. So, like, if you had loads of time, infinite money, infinite resources, and so on, you would ideally sit with the lawyer and say I'm doing this, is this correct? And you would get different answers from different lawyers, and you would try them all and see how it goes. But then sometimes, you just cannot deal with an ambiguity like you have no way to resolve it. In which case, you state the assumption that you think the ambiguity refers to and say that OK my working assumption is that this whatever statement means this and I'm going to work with that.

> So, you know so you will always get a certain amount of that. I suppose we do shy away from you know if we’re looking at a text and thinking, This is just really vague, you know, we’ll look at that and say well we’re going to get so much

> inconclusiveness, is this even the right strategy, let's find another, you know. Is there another way we can attack this, attack this problem you know. So, it's sort of interesting, for example, it was very useful for us that we found this corpora of other people who had annotated a load of privacy policies because we had been looking at that ourselves and saying Actually this is a really difficult job because actually those privacy policies are often professionally ambiguous.
> I: Okay.
> R: Because they're written by lawyers to try and give the most flexibility to the company and the, rather than making things too specific for the, you know, for the user and, you know, that's and again, that in itself has been studied you know. So there's papers we've referred to about the, you know the, if you like the professional ambiguity that's in there and again, you know, companies are sort of testing how, you know, how specific they can get in terms of what the market accepts and what people are happy to, and people aren't lawyers, they tend to sign anything, especially if what they're getting is free. You know, there's a, you know, I suppose a process of education but, you know, that's, you know, natural interest is not being more specific than they actually need to because they want to leave as many options open for secondary uses of that data as possible, you know, and it's something, even amongst privacy researchers, you know, you still get different views about

Legislation surrounding data use and appropriation can be deliberately ambiguous; what happens with your data:

> you know there's loads of people saying Oh there's no way I would, you know, like get these services where you give them their DNA and they come back with a little here's your profile…
> I: Yes, your ancestry.
> R: Your ancestors or, you know, your ethnic, you know, the ethnicity of your ancestors whatever which is great. But a lot of those things say And oh we'll hang onto your DNA for future use. That's not on, so if you choose what you look, hand over to Donald Trump in case he wants to decide whether I should be allowed into the US, you know you've got, you know like people should, you know, in principle they should be really, really wary of those sort of thing and I think some of the legislation doesn't even, you know, it just says you have to be specific but you can be specific in an ambiguous way.

To return to an excerpt discussed previously in section 4.2.2. regarding the perception of data definitions, "someone else's definition of what can be considered as personally identifiable may not be as extreme as someone else, in which case they have a problem if they're working on the same set of data." This has potentially serious ramifications if the legal text that defines these terms is itself professionally ambiguous or facilitates companies that have acquired your data being "specific in an ambiguous way" regarding what they intend to do with your data:

> So, someone else's definition of what can be considered as personally identifiable may not be as extreme as someone else, in which case they have a problem if they're working on the same set of data. So, if both of them are saying OK we are going to remove all identifiable pieces of information from this data, one person's

> dataset is going to be much larger than the other person's. So, sometimes when, like in that case, you're literally changing the definition of what is data. Like, this data may be a part of some other dataset, but how I derive this data from the bigger dataset itself defines what the smaller subset is. So that's, like, how do you define this relationship between different pieces of data? That needs to be clear. So, that's where you have gaps in communication. That's why you have legal text that should very clearly and unambiguously mention what it is and so on.

Contradictory data

Of the three kinds of difficult data investigated during the interviews, contradictory data elicited the least amount of responses. At least as a term, “contradictory data” did not connect with the researchers in the same immediate way that “ambiguous data” or “uncertain data” did. If data presented as contradictory, for the majority of those interviewed, it indicated the presence of an error, something that would be “washed away,” “ ignored, or like taken as noise”:

> you would assume if you have enough data that the kind of positive or the correct examples are more than the wrong examples. So if in your data is like “I love you” is the correct example, this is how it should be and there are like few instances that says “I hate you,” then they would be washed away by the algorithm anyway, they gonna be ignored, or like taken as noise.

> the algorithm would try to fit the most common examples first, and those kind of like outlaw examples they would be ignored if they are less. If you have like fifty percent and fifty percent, it's going to be very difficult, but usually in most cases like if you're doing like a cancer research and you have like few indicators that says if somebody has cancer or not and then somebody for example smokes and doesn't have cancer there is like you could basically assume that it doesn't effect, but then if you think about the large number of people then you could know that this factor is not really a decisive factor in this. So they usually, the data if was large enough it would usually find a pattern correctly and ignore those examples that are contradictory.

> Kind of…I mean, like I said again, it would happen, you would get contradictory data if in, like in the spreadsheets and stuff if, you know, one number is supposed to add up to another number and it doesn’t.

Again, context was advocated as a potential approach to dealing with contradictions in data:

> If you are analysing the context maybe you can realise that that's actually meant the opposite way.

> probably context also is giving you hints about whether or not something is contradictory or not. But out of the blue you cannot tell. If you are working with, if you have in the computer I love you and you have raw text, you are not having an audio or a video. How would the computer know that you actually mean I love you, or you're meaning, I hate you?

> But generally, we have a little bit again, a little bit more of context, so for example, if I have a sarcastic sentence with I love you, I would have, I love you, you wake me up at 5.00 in the morning to do your dishes. So, then you have like a negative situation expressed with positive words, so I love you is a positive word, but then you have a negative situation. So then if you put these two together, you have a sarcasm, which is a negative feeling. So then you have sarcasm, but you can only tell that if you have more context.

> again, depending on the context, you would either classify them as outliers and remove them as Oh this is contradictory data and we don't deal with contradictory data. Or, because I work with semantic web technologies, you have the open world assumption in which, it's this, like the assumption is that all knowledge can refer to all other knowledge, and just because something doesn't exist doesn't mean... just because something is not stated doesn't mean that it doesn't exist. So, for example, I love you and I hate you. So these are two different statements. So they could have been said at different times, they could have been said in different contexts. I love you for this. I hate you for that. So, I cannot make sense that this is contradictory unless you have some context to base these both on. So, if your context is I love you and I hate you then you'd say Oh wait a minute, you cannot feel both of them together, hence this is contradictory. So, you need a definition of what contradictory means in that particular context of data. And then you can say, and then you can classify data as being contradictory. So, for example, age. If you're taking the age number and you say that OK, give me your age and you say oh I'm 18 today and the next day you say Oh I'm 21. Then you say that, Wait a minute, this is contradictory. Like you can't have 18 on one day and 21 on the other day.

Contradictory data could even be indicative of malpractice:

> Yeah, like an error or like, you know, and for sure I mean, like what do you call it, there’s definitely companies and stuff which have two set of books, you know. Big companies which got done for fraud and stuff where they had one set of books that said one thing and then they had the real set of books that said another thing, but I wouldn’t call it contradictory data I mean, because you don’t have them next to each other. Something, it’s only contradictory if the single set of data that you are looking at is internally inconsistent. That’s the only time I would kind of think about it.

Contradictory data can also be data you do not want to hear about, or that differs with your worldview, ideologies, or beliefs:

> So, that's kind of like a story that you don't want to hear about, like, people are, but it's still part of the story and that's what I'm interested in is, like, you know, there's not just the one story. There's all these voices and all these things happened, and they all contributed to the bigger narrative. So it would be like being open to this contradictory, or dichotomous, or the untold stories.

While a number of the interviewees spoke of their algorithms washing away contradictions, others asserted that certain of this data cannot be removed:

> You cannot remove it. Maybe because it's not your dataset, maybe because it changes what the data is.
>
> Maybe I don't remove it but I put less priority to them. So the knowledge engineer will know that all, there is all this kind of data, this is not and the percentage of this appearance is not so high, so this, yes but I will not remove it either.

Uncertain data

Uncertain data, data whose "meaning is unresolved or unresolvable" was described by one interviewee as data without context:

> I think uncertain data is much more like, that's much more common. [...] Just, em, its meaning that is unresolved. I mean yeah without context, so without context it it's pretty much by default uncertain.

Again, the majority of the interviewees advocated the application of context as a means of resolving the uncertainty.

> For example like prevalence for TB is 2,800; in Kampala is will be considered as big portion but in Indonesia it is considered as small portion because the population of Indonesia is like ten times higher than the Kampala Uganda. So the 2,800 maybe treated in different ways.
>
> Put in relevant units. So for example, 2,800 is in units like 100,000 population in Indonesia but in Kampala Uganda it will be like maybe 10,000 population. So both are relevant units. So, the people who see it will not have two different meanings.

One interviewee advocated a statistical approach to dealing with uncertainty:

> We can have, for example, solution for the, like as, for example, we can just try to understand these, for example, circle meaning from getting feedback off different users, or different persons, then…predict something. You know that? Something like machine learning, the prediction is part of the life. For example, when I, if you bring it, okay, let me discuss better. For example, if 90 percent of the people say that this is a circle, it means that it's much predictable that it will be this circle. It will be the ball, for example. Yeah, I think that maybe we can solve this.

Uncertain data was pinpointed as data that could not be used.

> Is it good to have complex corpuses?
> R: To have a complex corpus, yes, to have an uncertain corpus, I wouldn't say so because I would, for me, if I have a misaligned corpus and I'm not certain that my data is good, for me it's not, it's not worth it because if I teach the machine, for example, that, let's say, football in my language, in Portuguese, is futebol. If I teach that in my aligned data I have that football translate as moon, lua, so I would just

have garbage in my machine and the machine's not going to do a great job. So if my data is uncertain, I am not going to use it probably.

Complex data is okay, like the complexity of identifying sarcasm in tweets, that's okay because you can have, how are you going to identify, for example, if it's sarcastic or if it's ironic. The complexity, it's there, but it's loads of work, especially if you want just to use the sarcastic ones, but it's still correct, let's say like this, and if you are uncertain, if your data is correct or not, then for me it's not useful.

if you don't know what it is, you don't know how it fits and you don't know how to use it, and that's what my whole thing is like What is it, how do we use it and how do we get people to engage with it? So, if you don't even know what it is, it really makes it hard to take it down the line of the next steps.

Uncertainty in data. I don't think it would ever be good, because that essentially means that you cannot use that data reliably, because it's uncertain. So, as much as possible you move towards structured information. Like, uncertain data is chaotic data, you don't know what's in it. So, you want to get rid of that uncertainty, you want everything structured proper, everything states something reliably. So, ideally you want to get rid of the uncertainty as much as possible.

Like, if you can remove the uncertainty, then it's fine. What happens if you cannot remove it? Do you just get another dataset? Do you ask someone else to do it? Like, sometimes you have an uncertainty, and you cannot use the data but you also cannot remove it, so it just sits somewhere for a while.

uncertain data is chaotic data, you don't know what's in it. So, you want to get rid of that uncertainty, you want everything structured proper, everything states something reliably.

Like, legally you, it's very difficult because legally you cannot say that This is uncertain data. Is this personally identifiable, is it not personally identifiable? I don't know, it's uncertain. If it's uncertain, then you're safer off identifying as personal information. So, uncertain means either, like, you need more processing, you need more data describing that data, it's yet incomplete, it's unusable. So, uncertain in what terms? Like, what is your basic assumption? So, it depends a lot on the environment that the data is being used or processed in. So, then uncertainty depends on that. Like, contradictory means that you essentially have at least two pieces of data which state two very opposite things. With uncertainty, it either means that oh this is incomplete, or it's expressing something that we cannot process yet.

This can, actually, yes, means something is uncertain. Unfortunately, it's unresolvable, and maybe those data can't use.

I can't say that it's a good things because when you have the data, but you don't understand the meaning of the data, it means that you can't use those data. It means that it's not useful for me. Even if I know that, for example, what's your weight, but I

> don’t know that this weight is, for example, is yours or is for someone else. How can I use those data?

4.2.5. Relationship between data and narrative.

How does data differ from narrative?

i) “We think in terms of stories, [...] we don’t think in terms of data.”

Data is not as readily conceptualised as narrative, or without a narrative of some sort to assist us in understanding the data. As the following interviewee puts it, “we think in terms of stories, [...] we don’t think in terms of data.” This then can be considered a key and integral difference between data and narrative: how readily one or the other can be digested by humans:

> I think, we, that’s what sort of data science and stuff, that’s why that is getting all popular, so popular now because we think in terms of stories, like as humans, we don’t think in terms of data**.** I mean, try and understand, that’s why graphs work so much better than spreadsheets because you just can’t look at it, it makes no intuitive sense. Stories make sense to people so it kind of makes sense that people would try and create a story about data, but you know anything that’s sort come across.

This interviewee notes, rather disparagingly, that this “think[ing] in terms of stories” has lead to the rise in popular science and “data science” non-fiction. But aside from the commercial realm of non-fiction popular science, this has implications that run to the very core of frontier research: so much emphasis is placed on data, and on the ethics of data use, data collection, data preservation, data sharing, but little or no emphasis is placed on the role played by narrative. Yet what was resoundingly clear throughout both the interviews and literature review I conducted for this facet of the KPLEX Project, was the centrality of narrative to the dissemination of all research, irrespective of one’s disciplinary background.

A huge amount of work and research goes into the production of these narrativisations of research data. This process, the process of getting to “The story I'm trying to tell at the end” is neatly outlined in the following interview extract:

> The story I'm trying to tell at the end [...] in order to get to that story, I had to go right back to gathering tweets, pulling out the tweets, and then analysing the words at part of speech level, testing out the tools that had been designed for grammatical texts on these tweets to see where it falls over. And then say Well this is why it's called noisy data, because it's all this stuff in here. And then analysing the ones that didn't fall into the usual categories and coming up with a way to categorise them. And labelling them all, and grouping them, and doing analysis on it. To then come to this conclusion that this I could say [...] So that's what all that work. So the data was much different to the story.

As this interviewee concludes, “So the data was much different to the story,” yet both are fundamentally connected, and the “story,” while different, is responsible for the dissemination of the data.

Before moving on, it is important to return to the key point from this section: “we think in terms of stories, [...] we don’t think in terms of data.” While we, as humans, “don’t think in terms of data,” by and large the technologies we employ to conduct and facilitate our research, the research that allows us to articulare our data “in terms of stories,” involves machines and softwares that “think” solely in terms of “data” and not stories. There is a disjunction then between the data, the software or technologies used to render this data intelligible, and the narrativised output of this process. The data is the facet of the project that humans often “can”t do,” especially when it concerns a large corpus of data, or when the data is multivalent. There is a balance then, between the analysis that “humans can’t do,” that computers or machines can do, and the act of narrativisation, which humans can do, but machines cannot do as effectively:

> Now I only had 15,000 tweets, it wasn't that big. It's hard as a human to say Here's a load of data and then you can make a conclusion just from reading through a couple of hundred of them [...] because you're making a grand statement just based on a few. So humans can't do that and that's why you need computers to find these

ii) Temporal differences and the role of sequence and structure

Narrative engages with temporality in a different way to data. It has a temporal dimension that data do not. As the following interview excerpt makes clear, “Narrative also has some process [...] it has a certain timeline. Not in strictly in parts of time, but it’s like it has a certain order”:

> Data itself is just structured information. Narrative also has some process or, like, narrative comes from, like, story basically. Which means that it has a certain timeline. Not in strictly in parts of time, but it’s like it has a certain order saying that Oh this was before, this is what happened, this is what happened next. So, you have a certain order of things.

Many of the interviewees observed that narrative emerges *after* data, as point that was also discussed in the previous section wherein an interviewee noted that “The story I'm trying to tell at the end” only came to fruition following a lengthy engagement with the data:

> I think the narrative comes after an interpretation of data. So, it's like you look into the data, you give yourself some kind of idea about what's going on. So, it's like yeah you're adding basically, an interpretation, and that of course can be false.
>
> once I've surveyed all of that data and seen what’s out there, kind of, a narrative will emerge, trends will emerge, there'll be big pockets of "this is important", "this is important", "this is important" and then using the data to supplement it, and create, kind of, the story.

> there is a temporal difference because there is a moment in which the data is being collected and then usually the narrative comes after, sometimes even years.

(Data/ structured information) Seems to be reaching here towards the idea of narrative as the act of adding structure (story-like), also narrative as description or research process.

> Data itself is just structured information. Narrative also has some process or, like, narrative comes from, like, story basically. Which means that it has a certain timeline. Not in strictly in parts of time, but it's like it has a certain order saying that Oh this was before, this is what happened, this is what happened next. So, you have a certain order of things. So, you can have data at different stages and then the narrative could be, like, this was how we processed it.

Across the board of interviewees, many of whom noted the role of temporality in differentiating narrative and data, it was regularly observed that narrative should not precede the data and that a researcher or scientist should not "pick the data to suit their story.":

> some people will have an idea and go I wanna prove this and if they're a bad scientist they will pick the data to suit their story. If they're a good scientist they'll take all the data that's out there and then come up with ways of analysing it and hopefully report good and bad results.

Granted, how one approaches data at the onset of a study "depends on what exactly what you want to look, for me it also means what story do you want to tell, or do you want, the story that you want to write." In other words, on your hypothesis. This then is another instance of the centrality and importance of ethics and researcher integrity to the narrativisation of data :

> I think, generally when I said to you in the beginning that your data depends on what exactly what you want to look, for me it also means what story do you want to tell, or do you want, the story that you want to write. So, yes I think the data is very related to narratives, but you know, sometimes it's, it's a little bit hard because sometimes it doesn't matter how many tests you do in you data, it is not the narrative you want, but still is a narrative, doesn't matter the result that you're going to get from that or not. What it just makes me a little bit cautious, is that depending on how much of data you have or how, how careful were you to handle your data, or to explain your data, how much of a narrative can you actually tell on the top of that. How much conclusions can you take, apart from that. So, for example, if it says in the first one, the patient ingested a drug and their condition improved. Are you sure it was just that drug that he was taking or not? Like, is there anything behind that that made, maybe? So, you know, plants that are not watered will die. True, but plants that are watered, they will also die. You know, so I think yeah there is a narrative, but I always try to be very careful in explaining how I got to that one, and it's very, and because I work with human evaluation, we never have that many humans helping us out. So, if I have a task maximum that I generally can get, it's like two, three translators, maybe five translators, when I can get five translators, I'm in heaven, I'm like wow, so how much can I actually claim that this translation is great because three of my five

> translators says it was great, but two of them say it was really not that great. So, you know, so we also have like a bunch of metrics for annotator agreement, and people don't agree with each other, they don't want agree with themselves, you know like I said if I give you a sentence to translate now and in one week I give it again, your translation, you're going to look and say, Oh no, this is not good, like I'm going to change this word for this one. So, how much can you actually tell what it is with the type of data you were. So there is a narrative, like I can always tell a narrative with my data, but how much of that should I claim that it is what it is.

Data is easier to narrativise once it has been structured or sequenced. The act of sequencing or structuring the data seems to initiate the temporalisation process that facilitates the narration of the data, or as one interviewee puts it "building that sort of temporal element into what we're doing because we think that is a little bit easier to perhaps convey":

> so part of what we're doing is, you know, trying to meld that sort of very rule based approached with something that has a little more, and we talk about it more as a, process flow. That says okay, there are certain things that are always going to happen in a certain order. Okay, so this isn't quite yet a narrative but it's building that sort of temporal element into what we're doing because we think that is a little bit easier to perhaps convey and then it comes to how would you actually convey that.
>
> Well I think it would, like from a technical point of view, you'd have a bit of a narrative in that you'd have, you know there'd be a sequence of things that would happen and an outcome.
>
> numerical or categorically data like spreadsheets and just creating a text summary or just description of the data.

The narrative is situated in a space that succeeds the processing and sequencing of the data, it is a narrativised account of the "outcome" of the process of datafication and data analysis.

Elsewhere, narrative was defined as the relating of cause and effect, and as such then as something that recounts a series of events, which explains why data that has been serialised or sequenced into a pattern is more readily narrativised:

> So, in these stories like if it's, so it's a cause and effect, are present here. Cause and effects series of events.

Furthermore, one of the key characteristics of narrative, of story and the story one tells about ones data, is that the narrative must be connected in a logical manner:

> Story has a, I can put five lines together but they are not interconnected. They has to be interconnected and they has to have to have a coupling between them, and which we can logically deduce in our head. Because if I say *The plants that are not watered will die. We have photos of all the guests.* That is not a story

This same interviewee, in making their point regarding the necessity of connectedness to story functionality, notes the distinction between story and dream: the connections in a story or narrative have to be visible, and acceptable or believable. In a conscious narrative/ story, what is acceptable is grossly different to what is acceptable in a dream:

> So when we merge them together, we need to have a thread [...] So that's why I said the sequence of events and we are like really mixed up where, like… Stories are where you go, even for children's story. It's still feasible to align with another story we have, our life story, it's documentary if you consider documentary. It's something, it's believable, so it's happens, a series of connected events. Comes to, now if you consider the age slowing down, when you tell a story to a kid, those stories does not have so much meaning. They have connecteds. They still have dragons which is not exist, doesn't exist or ghost stories we create. Now come down to dreams, where you combine two things which never exist. So we find out the series of events, that the connection depending on what you are talking it's a child a children's story, or it's a dream, or it's a documentary, or it's a romantic comedy. So, all different genres of movies, because there are timelines. So in that case we find out the connection will be, but they're all data all stories, but the difference will be is it understandable by a human mind or not. Because there are, a friend of mine said a story that he tied cabbage on a jeep in his dream. So and we find out, Okay, yes you can tie a cabbage, you can still make them, you can still visualise them. So that's needed because otherwise these two lines, we cannot visualise them, how that could jump from. So jumps should not be too much, and that's where it breaks the story.

Narrative is presented as a connected series of events, events whose connectedness is believable. Data, in contrast, when it is what the following interviewee describes as "loose," "doesn't really have much of a story":

> So from what I've learned, I think narrative is a more loose description of what, you know, is a connected series of events and story tends to have more specific components in it, like it needs to have A, B, C to actually be considered a story. So, data might be the same thing where maybe there's loose data that doesn't really have much of a story. Maybe it has a bit of narrative to it, but then there might be other data that it's very clear that it has a solid narrative or a solid story. It's just a matter of what's there and how whoever's interpreting it interprets it.

Data by itself is not or will not have a narrative:

> yes, data can have a narrative if you, like data by itself will not have a narrative. You would need some process or some additional metadata that describes the narrative itself. So, I will strictly keep to these examples. So, you have the patient ingested the drug and their condition improves. So, you have a bit of data that said OK yeah, the drug, the patient was administered the drug, their condition significantly improved. So you have a measure of OK, this was the condition before, this was the condition after. And then your inference becomes the drug helped alleviate the condition. So, you're generating new data from previous data. So this process of deriving data from data can be a narrative.

This interviewee notes that metadata (itself a form of context) can be considered a narrativisation of the data, given that is is “data” that has been derived from “data.”

A contradictory point can be made here, data itself does not necessarily have to be sequential, it does not have to be structured. As the following interviewee points out, “data is not sequential,” data is “unstructured material” and there “doesn't have to be like there’s any kind of causality between it.” Considering the extent to which data, in some instances, may have to be sequenced or structured in order to be represented or rendered narrativisable, it is possible to argue that the addition of structure or sequence to data, any transformation that renders it sequential, is an act of proto-narrativisation:

> I think that data is not, it doesn't have to be like sequential or it doesn't have to be like there’s any kind of causality between it. It's very, I think of it as a very unstructured material, while like a narrative is more like it feels a bit more like there’s a story and sequence between.

A key point we can take from this then is that data representation, insofar as representation involves the imposition of a sequence, structure, or order, is an act of narrativisation.

What is the relationship between data and narrative?

There are a multitude of potential avenues through which one can narrativise data, as outlined in the following interview extract:

> the way in which the narrative and the data and the sort of the model, especially models, the way that all kind of works together, that’s eh, there can be just so many different ways in which it is constructed. I think, we, that’s what sort of data science and stuff, that’s why that is getting all popular, so popular now because we think in terms of stories, like as humans, we don’t think in terms of data. I mean, try and understand, that’s why graphs work so much better than spreadsheets because you just can’t look at it, it makes no intuitive sense. Stories make sense to people so it kind of makes sense that people would try and create a story about data.

As this interviewee makes clear, people “don’t think in terms of data.” As a consequence then, narrative has an integral role to play as the mediator between data and its audience. The importance and pivotal role played by narrative in the explication of data deserves further recognition, and will be returned to in Section 5’s Discussion and Conclusion.

Disciplinary background may influence the narrative one tells about data, with one interviewee observing “You will be your own bias” and noting how one’s approach to research is influenced by the brief one receives or the project objectives:

> One of the, when I was studying I did this computational linguistics program but I also did translation and one of the exercises we had was actually, they gave us a text to translate and they gave us the message from the client. And half of the class had the message was coming from Greenpeace or an NGO that was trying to protect the animals. It was about a church that was having a problem with pigeons.

> And they, and then there was, the priests wanted to get rid of the pigeons because the structure of the church was in danger, and then the NGO wanted to protect the pigeons because they are animals. And then half of the class was given the email from the NGO, the other half was given the email from the priests or whatever, or the, I don't know, it was the priests or the town hall or whatever. But they wanted to get rid of the pigeons, right? And it was funny how then you subconsciously just because of the type of client you have, you choose different ways. And then we came up with two different perfectly valid translations that were hinting towards one or the other direction because when you were at the situation where you had to choose, you chose the one that were, that was given by your client. You were thinking of the client that was receiving the translation. When you don't have that instruction then it will be your mind. You will be your own bias.

An identical sentiment was expressed by another interviewee, who noted that one of the factors you "add to the data in order to get the narrative" can reflect your disciplinary background or personal beliefs, conscious or unconscious:

> And there is also something that you add to the data in order to get the narrative, and that's, I don't know, you are…this actually reflects your own beliefs sometimes. You are sometimes even in your bias, things that you are not even aware of.

Indeed, one interviewee noted that narrative is understood very differently depending on your disciplinary background:

> it's quite interesting to see how, you know, people in arts and literature approach narrative, and then how computing scientists approach narrative, because they have very different approaches, very different definitions of these two things. And so it's a matter of reading both sides of the story and trying to find a medium where both sides could understand what I would be talking about as, kind of, that middle ground. So, that's been interesting to see how they do that because the narrative people have, you know, they've done literature for years and they know it inside and out and then, the computing scientists are taking these technologies to be like, OK well we can use this technology to, like, change narrative and kind of mould it in a different way for the digital medium now.

In the evolving relationship between data and narrative, we see the emergence of new disciplines and new ways of conceiving of narrative:

> lots of narratives traditionally told linear, beginning middle end that's it. It doesn't change, it's published, it's set. And my whole kind of philosophy/background is No, the digital medium now, it's two way. There's not just the author, there's also the reader, and the reader has knowledge and they want to participate and they want to interact with something. They want to contribute. So you can change the narrative and build on it with the digital medium.

> we now have all these digital mediums. When is it the best to use which one and how, and how does it add value rather than take away or be an interruption in the narrative experience?

Furthermore, with these new digital media facilitating an increased level of participation, the potential for narratives emerging from extent data is heighten as more and more people interact with and draw narratives from the data:

> So you can, kind of, create a prototype story, for example, and then when the users interact with it, or other people, they can contribute to it or supplement the facts

One interviewee identified the relation between data and narrative as symbiotic, noting that "you can find a narrative from data or you can use data to make a narrative":

> I: what do you think is the relation between data and narrative?
> R: Symbiotic? I think you can use data to make narrative and narrative can inform the data, like vice versa. So, you can find a narrative from data or you can use data to make a narrative. So, I think they can go hand in hand one way or another and maybe, you know, cross a bit. So, for mine, I'll be looking at data and seeing what narrative can emerge from it, rather than me projecting a narrative onto the data, but also I'll have, kind of, an overarching narrative in the back of my mind. So, I won't be going in hunting for that but if nothing emerges, I might be able to come with that and create something out of it so it's kind of like a balance.

Again, this gives rise to the issue of ethics, and at the need to refrain from "projecting a narrative onto the data."

Another interviewee described the relationship as one of "influence," with narrative influencing how we approach the data, and data conversely influencing the kind of narrative we tell about the data:

> Well it will always influence, I guess. The way, the way you describe your data, it's in a way, the narrative; I mean it's usually conditioned by your research. So, if someone else was doing research on the same data as I do, probably they would describe it in a different way. That's not necessarily, but sometimes may lead to problems. Because, especially when you are trying to answer a research questions and what would happen in the end was that sometimes you will describe things in a way that seems to direct the thoughts of the reader towards the idea that you have in your brain. We're using rhetoric, we're using the narrative to guide the reader in our experiment but it's also about how we describe the data that will... So, yeah, hmmm. I think there is an influence there.

Not all interviewees were as forthcoming in seeking to establish or validate a concrete relationship between data and narrative, with one researcher arguing that there is no relationship between the two, there is only "lies and statistics":

> No. There's lies, lies and statistics or, you know like that one. Is there a narrative, I mean for sure, you tell a story and then you support it with text or with not with text,

> with data. I mean the story itself can come from the data [...] But these are like, these are kind of modelling issues, somewhat, I would say lots of people attended my wedding we have photos of all the guests. That more of like a, Do you have photos of guests? Do you have just photos of people? You know, what's the connection there? I can see some of these are, they are different kinds of correlations, and obviously some of them you know plants that are not watered will die. Some of them are like self-evident and definitely true.

The above account is contradictory at points, for instance in the switch from arguing against their being a relationship between narrative and data, followed almost immediately by their stating "Is there a narrative, I mean for sure, you tell a story and then you support it with text or with not with text, with data. I mean the story itself can come from the data." The interviewee highlights this distinction of this relationship as a "modelling" issue, which makes it representational and founded on the ethics and integrity of the researcher to ensure that the connectivities being asserted by either the narrative or the data (or both) are "self-evident and definitely true."

This same interviewee also argued that if "data is truthless," narrative is "completely fake":

> I mean, narrative is, you know, it's like what was the earlier thing that was mentioned where like data is truthless, you know. I would say that narrative is much more like, that's completely fake. Like, any sort of model or meaning or you know representation or any of these kind of stories that we come up with, they're not an accurate representation of reality, for sure. And as much as possible, people try to use data to back it up, to show that their narrative, their representation of the world is correct.

This is a reversal of Rosenthal's argument that data is used to backup (fictional) narrative. The prevalence of this confusion regarding the relationship between data and narrative points towards a need to clarify the functionality of narrative as it is employed as a tool by researchers to communicate their data, particularly as it is understood among researchers working within the computer sciences.

> Narratives are mostly not false, but they are all made up, but you do need, sort of data or something to back them up, or at least that's my feeling on it I guess.

Elsewhere, one interviewee argued that narrative is "data based on data":

> Well, like all narratives, if they involve data, then the narrative itself will be based on the data. Therefore, it is new data so it is, like, narratives will be data based on data.

When asked to elaborate, the interviewee clarified that, as data on data, narrative was metadata, or reasoned/ interenced data based off ot the original data:

> So, it can be something that describes the data in which case you're actually talking about metadata, or it is generating new data from some data, in which case it's reasoning or inferencing.

What is the function of narrative

i) To explicate and contextualise data.

Our interviewees elaborated on what they perceived to be the function of narrative, with several observing that narrative was necessary to explicate data. For one interviewee, data cannot be understood without the assistance of narrative, with narrative providing or supplementing the "meaning of the data": "because if we only give them like graph or tabular, people will not know the meaning of the data." For another interviewee, the role of narrative was to "contextualise" data and indicate the functionality of the data: "I probably have to contextualise and say what can you do with it."

Elsewhere, others picked up on the idea of narrative as description that allows you to "learn something" from the data, of narrative as a way "describe that [data] in a more like informative way":

> I think they would like typically you would have a story. Like you would be able to learn something from the data and then um, like being able to tell this in a scenario, I think it's possible there is kind of um, like it provides you with a lot of insights and I think it's a good way to, kind of, describe that in a more like informative way.

Narrative then can be considered a form of data representation that "translate[s]" the data into a more accessible or understandable format, "translate into something that you know, anybody could look at":

> I think narrative helps people communicate what they found because not everyone can understand data on its own, so you kind of need the narrative to help explain it.

One interviewee whose research is interdisciplinary and straddling computer science, the digital humanities and cultural heritage, spoke of narrative as something that could be altered depending on the perceived audience, and as narrative as something that was treated or approached differently depending on ones disciplinary background:

> it's quite interesting to see how, you know, people in arts and literature approach narrative, and then how computing scientists approach narrative, because they have very different approaches, very different definitions of these two things. And so it's a matter of reading both sides of the story and trying to find a medium where both sides could understand what I would be talking about as, kind of, that middle ground. So, that's been interesting to see how they do that because the narrative people have, you know, they've done literature for years and they know it inside and out and then, the computing scientists are taking these technologies to be like, OK well we can use this technology to, like, change narrative and kind of mould it in a different way for the digital medium now.

ii) To add "structure" to data.

Narrative was presented by several interviewees as something you could use to structure your data, as a framework into which you inserted your data, and as a means of mediating data access; all of which are complimentary extensions of the previous point regarding narrative as a means of explicating and contextualising data:

> I hope to develop a framework of OK you want to create this thing, these are the things you need to think about, and this is how you can do it. So that I can help future people who want to create a cultural heritage narrative, or something like that. They can, they kinda have the framework and the steps and, like, the formula of how to do it in a way that will achieve their goal.

In a digital environment, this narrativised data or "what lies beneath" can be mediated by computer interfaces, which related back to issues regarding digital skeuomorphism and the transition from analogue to digital environments touched on throughout the Literature Review: "So, the narrative that I want to tell is mediated by a website where people can actively find out what lies beneath."

In emerging digital environments, however, rather than being tied or restricted to these skeuomorphic replicants of analogue narratives, narrative (and the data it narrativises) can be presented in more non-linear manner:

> So if someone wanted to read a certain narrative, they could just follow the digital path to that narrative they want to read because they can have choice. But if they're interested in seeing some of the side branches and, you know, different narratives, they can also veer off and go read about that and, kind of, get back on track if they want. So, that's the beauty of the non-linear narrative structure.

Again, however, even if this narrative is "non-linear" it nevertheless imparts a structure onto the data, and this structure makes the data more accessible to human agents. The popularity of the non-fiction genre of popular science was used by one interviewee as an example:

> It really, it just, like, I would say absolutely yes, but the way in which the narrative and the data and the sort of the model, especially models, the way that all kind of works together, that's eh, there can be just so many different ways in which it is constructed. I think, we, that's what sort of data science and stuff, that's why that is getting all popular, so popular now because we think in terms of stories, like as humans, we don't think in terms of data. I mean, try and understand, that's why graphs work so much better than spreadsheets because you just can't look at it, it makes no intuitive sense. Stories make sense to people so it kind of makes sense that people would try and create a story about data, but you know anything that's sort come across.

What this excerpt elucidates is that narrative need not be textual, that, as the interviewee points out, "graphs work so much better than spreadsheets because you just can't look at it,

it makes no intuitive sense.” Again, this ties in with section 4.2.5.i) regarding the function of narrative: to explicate and contextualise the data.

This is especially useful when data is multivalent and coming from a variety of different sources and different formats, with the subsequent narrative is an amalgamation of these formats that will combine through narrative “answer your research questions”:

> you could even have keystrokes, like recording exactly what they did in the keyboard. And if you have an eye tracker like we do in ———, you could even record where their eyes are going, to identify the words that are more problematic. So then you have like different dimensions of data. There is text but then there you have the time, and you have the fixations of the eyes, and all that together has to be merged somehow to, to answer your research questions.

iii) To make data accessible.

By making data accessible and explaining data, narrative has the potential to democratise data access and data interpretation. One interviewee expressed the hope that narrative could facilitate a more involved participation with data:

> I would hope there would be at least, you know, a branch or a path that they could follow, and if there were other people who had stories or materials to contribute to it they could. And then it could become a whole living branch of the story.

In making data “understandable or reachable,” narrative can be seen to democratise data access:

> it kind of makes it understandable or reachable for someone who might not be an expert or understand what it means

This naturally can also be exploited, with one interviewee cautioning that in instigating causality, narrative may also distract from “hidden variables”:

> So it's like, for every statisticians you should be always aware of causality, casuality in terms of, there is a strong difference between these two. So in planning your experiment, you should always keep this in mind. And sometimes you do some experiments, as I mentioned before, there are hidden variables. Sometimes there are hidden variables that you are not even aware of. So they can actually blind you when you build this kind of narrative because maybe you are omitting...

This capacity of narrative to gloss, distract from, obscure or obfuscate means that there are ethical dimensions to narrativising data, or using narrative to explicate data. Because narrative can influence how we interpret data, it can also misguide or encourage you to interpret “things in a different way”:

> there is also the problem of us sometimes interpreting things wrongly because you're taking into account other things. Or the narrative misguided you and then you interpreted things in a different way.

> You know so actually how do you turn it into a bit more of a story, again on that basis that people react well to stories, you know, they understand stories with characters in it and especially when you know because obviously data protection is a bit dry, you know and part of the problem sometimes is, it's, people are very quick to say oh these guys, Google is bad, or Facebook is bad you know and often it's not necessarily people bad it's just they're not looking after your data and they're not thinking about the badness may come about because you haven't protected something potentially and then somebody else takes advantage of it later on. So it's, you have all these contingencies about what might happen.

iv) Ethics of narrativising data

For one researcher, narrative is the "ethical" provision of context to data:

> The narrative, it for me, is like giving the context of the data so that people will not mislead, misunderstand the data, the data will be like tabular or graph or yes.

Narrative shapes how one perceives data, and this is particularly important when it comes to its interaction with non-experts, with people who can only rely on the narrative, because they cannot themselves read or interpret the data:

> The narrative will shape the, the perception of the data to be read by other people who are not in our background.

This leads inevitably to the issue of ethics and misuse of narrative, about which several interviewees elaborated on. It should be, as one of the interviewees points out, that "the data influences the narrative":

> it should be, in my opinion, that the data influences the narrative, because you are doing research on data and then you are coming up with results out of that data. It could be that a researcher is more, is biased towards the research questions and then uses the data to justify what they were looking for.

However, should the researcher be biased, the narrative can be altered and manipulated to misrepresent the data.

The temporal distinction between narrative and data, with narrative emerging after the study or analysis of data, was observed by several of the interviewees as a core distinction between data and narrative, discussed previously:

> The narrative always, is something that you always do, because when you're going to write some conclusion about, so you are definitely drawing some kind of ah... There is an hypothesis that you had at the beginning, you have done something on

> the data and you actually have like validated your hypothesis, or not, depending on the measure that you have used and you got. And so you build actually a story from this, like okay my hypothesis is that the data tell me this story and then I can tell you that this hypothesis is true and whatever. So, there is always a narrative. That's sometimes dangerous because, depending on how you do your experiment, sometimes you can find some cause effect relationship that are not true.

Further still, even when used in an ethical manner, one interviewee stressed the importance of stressing the potentiality for other factors to influence and alter the outcome of one's research, and in particular for researchers to admit when their outcomes were unexpected, undesirable, or incorrect:

> so, you just have to be honest in the end and be ethical and say like, Yeah I was expecting that, I was wrong. Yeah, the patient got, you know, improved his condition taking the drug, but we cannot forget that there is this and this and that, so, that's what I think at least.

It is down to the ethics of the researcher to admit when their data "is not telling":

> The narrative is there, but you know, I think that that's our part, that's the ethical part of being a researcher, you know, it's like, just because you want the results to be like that, it doesn't, you know, sometimes your data is not telling you know.

> Probably it's like how open minded they are and not being afraid to fail. Because if you go in with a hypothesis, and the testing and the data shows that you're really wrong, my supervisor said that in the arts and humanities it's OK because you've shown something, like, you've proven, you've disproven yourself or you've asked a question and shown that it's not, so you've done something of use and value. But I think in computing science perhaps, or another, like, a hard science, maybe that is seen as failure rather than, you know, you've made a contribution in the end.

There is a responsibility at a disciplinary level to encourage ethical behaviour re data and narrative, with "failure" in terms of unsatisfactory results in the hard sciences and computer sciences presented as more frowned upon or taboo than "failure" or unsuccessful hypotheses in the arts and humanities:

> Probably it's like how open minded they are and not being afraid to fail. Because if you go in with a hypothesis, and the testing and the data shows that you're really wrong, my supervisor said that in the arts and humanities it's OK because you've shown something, like, you've proven, you've disproven yourself or you've asked a question and shown that it's not, so you've done something of use and value. But I think in computing science perhaps, or another, like, a hard science, maybe that is seen as failure rather than, you know, you've made a contribution in the end.

Lastly, an additional ethical questions to be addressed in terms of the relationship between data and narrative, and the function of narrative in emergent digital mediums in particular, is

the emergence (or indeed the suppression) of counternarratives, that people may or may not want to hear about:

> kind of like a story that you don't want to hear about, like, people are, but it's still part of the story and that's what I'm interested in is, like, you know, there's not just the one story. There's all these voices and all these things happened, and they all contributed to the bigger narrative. So it would be like being open to this contradictory, or dichotomous, or the untold stories.

This was discussed in section 2.1.6 of the literature review and Presner's discussion of "lack of indexable content [...] or even content that the indexer doesn't want to draw attention to (such as racist sentiments against Hispanics, for example, in one testimony)."[331]

v) Narrative and uncertainty

There is a particular issue in terms of how we deal with uncertainty in narrative, or how or whether we should narrativise uncertain data. In the following passage, the interviewee identifies the researcher as the key figure in deciding what is important (narrativised) and what is left out:

> I think it would, for my purposes, depend on how important it was. So, for something that seemed very important to a narrative and should be included, it probably would get included but with the question, like, What is this? and maybe that would be a way to get people to interact with the content and spark conversations and ideas and get people interested in cultural heritage because there's this thing no one knows about and people love mysteries so, that would really cause user engagement. But if it was something that is so, like, uncertain and you don't know how it's relevant or not, then maybe it would just get, you know, We have this material and it may come up at some point. Maybe it'll get forgotten, but not included at the time. It would depend on what it is and how important you guess it could be.

For one interviewee, "uncertain data" was data without context, with the context serving to supplement a narrative to make sense of the data:

> It's hard to figure out what an uncertain piece of data might be at this point, depending on what it would be. But, it could be, just like, let's say an image and you've no idea what the image is about, or you don't know if it's important to a story. So if you couldn't create a narrative out of it, it doesn't really have a place in the narrative, but, you could crowd-source it and see, like, What do other people think? Maybe people are working on digital software and they can like figure out what the image is and stuff so…

[331] Presner, in ibid.

5. Discussion and Conclusion

Interviewees confirmed the findings both of the literature review and data mining exercise: people define data differently, people interpret data differently, and 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 while they would indicate that Rosenberg's concept of data as "rhetorical" still holds true, the sheer scale and variance has significant more impact than what it is possible to convey with the single term "rhetoric." 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, several interviewees referred to data 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.*

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 that have established clear communication practices regarding the working definitions specific to that project, or researchers on a small team who are familiar with each other, and have developed a more intuitive understanding of each others research habits, methods, and preferences.

The interviews suggest that the phenomenon of overdetermined terminology and its capacity to influence research and innovation can be counteracted on small scale by agreeing on set definitions of terminology within set projects or working groups. For larger scale projects, agreeing on set definitions, remaining consistent in their usage, and flagging these instances of subjectivity and interpretation if and when they occur, would be particularly beneficial. Furthermore, transparency regarding the transformations the data have undergone, acts of pre-processing, processing, and cleaning must be flagged for the attention of other potential users of this data. Accountability regarding both the provenance of a given dataset, and of the researchers involved in the collection and curation of that data, therefore emerges as a key recommendation. There is a clear need to revisit the DIKW hierarchy and the purpose (if any) that is serves, as the segregation between terms that it implies is not only inaccurate and outdated, but misleading. Is its function purely theoretical, and for instructive purposes only? If so, it needs revising, because it does not reflect how computer scientists talk about data in the field.

Terminological comorbidity and counter-narratives to the DIKW hierarchy was observed throughout, particularly between the terms data and information, with "data" being referred to as something that contains "information." Further still, the same term is used to refer to different data within the one overarching research project (data stream, data cluster, original data, evolved data, evolving data), and to data as a synecdochical term covering both the whole and the part. The same term was also often used to refer both to specific data/ datasets & more general data / datasets, and to refer to data from different phases of the

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. Again this presents as a problem that is surmountable on small scale levels, or even at disciplinary levels where common-sense, familiarity, and a shared disciplinary background renders the interlocutors fluent in and familiar with each others terminologies and research methods, but not on larger scale interdisciplinary levels, or in environments where the backgrounds or motivations of researchers/ participants are not necessarily known or trusted. 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 detailed provenance, greater transparency regarding the transformations applied to data, and increased accountability on the part of those responsible for the curation of data would help counteract this distrust and dissuade insular behaviour. Researcher accountability was highlighted not only as a key factor in assessing the validity or reliability of data, but 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 CPD regarding Best Practice apropos data curation, processing, and analysis are additional recommendations. Regarding complexity in data, context and disambiguation emerged as the two key methods that, when combined, serve to effectively tackle ambiguity or uncertainty in data. This was repeated almost unanimously across all interviews, with many of the interviewees outlining the already well-established methodologies for dealing with ambiguity etc that are specific to their disciplines. The interpretation of data has therefore overwhelmingly been cited as researcher and context dependent, with the processes of disambiguation (itself another example of context dependency) cited as a key method for identifying and dealing with uncertainty and complexity apropos data. This gives us three factors—context, disambiguation, and researcher due diligence in the form of the provision of adequate training in and adherence to Best Practice and the Ethics of data collection, cleaning, and curation.

Moving on to data itself and WP2’s core task of moving towards a reconceptualisation of data: Data definitions, such as the taxonomy of data definitions outlined throughout the literature review, are only partially effective in terms of their ability to define data. To say that data is “rhetoric,” for example, may indeed capture the protean nature of data, but it is of little practical help. It also places emphasis on the interpreter of the data, and the subjective nature of our interactions with data, as opposed to the data itself, insofar as it is possible to outline its contents or constituents in an objective manner. To categorise data according to how it is processed, such as in the NASA EOS DIS, allows for greater transparency regarding data transformations, but again as outlined in the literature review, these categorisations only extend to the point where the data is acquired and used for further research: so that highly processed, Level 4 data becomes “raw” data. This focus on the processing rather than the contents seems to be effective in the field of Earth sciences, because a description of contents is comparatively easier to do than with more complex cultural data, wherein even the act of describing the entity risks misrepresenting it. This contrast between the relative consensus surrounding maths and earth science data versus humanities or cultural data was noted in the interviews:

> *Sometimes people can't agree, so controversial issues don't get agreed and get marked as things haven't got a consensus around it. You also see for example that you know it doesn't try and like Wikipedia doesn't do much translation okay, they say well they do it in certain like physical sciences where there is an objective agreement on, you know, maths or anatomy or something like that but loads of other things they say well no, different countries and cultures should find their own version of Gallipoli or you know, World War II or something like that because they will have their own, you know, genuine perspectives on it you know and even within a single language, people are going to disagree about it and the process of writing a Wikipedia page is almost a process of deciding what you're going to leave out because you can't agree on it.*

Focusing on the processing of the data rather than its contents is a necessary *facet* in this move towards a reconceptualisation of data, but because of issues relating to subjectivity in relation to contents description, it can only ever be a facet and should not be the end result in and of itself.

Of greater use is the idea of data as input. From a terminological perspective, the concept of data as "input" perhaps does not align with Druker's claim that data is "given" whereas is "capta" is taken, because there is the potential for ambiguity regarding whether or not the act of inputting results is data or capta. For example, the *capta* may be "taken" from an input environment, curated and "given" as *data* in the form of a dataset. For the curator, the inputted data naturally leads to an understanding of "data as input," whereas for the user of data, who is subsequently *given* this dataset.

Building on this, it would be fair to say that data are content. A tripartite focus on context, content and changes (in the form of how the material has been treated) allows for greater transparency regarding what data are, how (in)complete a dataset may be, how reliable it is, and who it could be useful for. Our understanding of data depends on what the data literally are on a project by project basis. And this subsequently dictates how the data are conceived of and spoken about. Content focused explications of the material, together with absolute transparency regarding the provenance not just of the dataset, but of the processing, pre-processing and/ or cleaning the data have undergone, is a necessity. This is already considered Best Practice in certain disciplined, such as computational linguistics, but does not seem to be as diligently adhered to in the computer sciences, or is done only idiosyncratically or inconsistently. One can imagine that such practices will be similarly inconsistent in disciplines with emergent or marginal interests in the digital, such as the branches of literature, linguistics and history (to name but these) that are engaging with the digital via the platform of the Digital Humanities. It is necessary to develop clear cross-disciplinary policy to guide researcher towards Best Practice regarding their data practices.

Interviewees rejected many of the definitions considered most authoritative or to hold most traction among the DH community, referred to them as "anthropomorphised," or simply stated that they were impractical, overly theoretical, or philosophical. Those that were more accommodating or positive towards the definitions still noted their philosophical or abstracted nature, often remarking that they had never thought of data in that way before. What data are can only be partially defined by all-encompassing statements of elucidations; particularly

when that data is cultural. Rather than focusing on identifying an ur-definition of data, our capacity to describe or delineate data in a way that does not delimit the interpretative potential (latent or otherwise) of the material being described, the curation of cultural data for the purpose of its inclusion in digital collections (etc) could be further granularised by making it content dependent and focusing on data provenance and transparency regarding processing.

There is a sense that humanities scholars are unprepared/ ill equipped/ unfamiliar with the technologies and methodologies used in the DH, technologies that have been adapted from science and computer technology. They are thus more likely to blame/ ignore/ discard/ dismiss the epistemological fallout brought on by the digitizing process (etc.) than adapt the processes to suit the materials being represented or digitized. Alternatively, and this is the point noted by Borgman, they may blindly adopt these processes, without considering the epistemological implications of doing so, or without considering that these techniques, when inducted into a humanities environment, must necessarily be examined from an epistemological perspective and become more fluent parlance and more integrated into epistemological studies of humanities research. This is further complicated by the fact that within computer science the act of “borrowing” or “repurposing” software is itself common practice

> You know, there’s always this sort of contextual model, but it depends on your viewpoint and I think the lesson learned from the computer scientists was you can never resolve that because people genuinely have different viewpoints okay, and trying to make everybody conform to the same model isn’t helpful you know because they won’t agree and they’ll go off and do their own separate ones, you know, which happens all the time.

Legislation surrounding what data are in relation to personal data and personally identifiable information needs to not be ambiguous or designed to facilitate ambiguity in relation to the sharing of personal data without fully informed consent.

Moving on, of particular interest is the ongoing debate over the relationship between data and narrative. This is a facet that has received significant scholarly attention, to date, but the research conducted by WP2 has shown that the extant field of scholarly inquiry into data and narrative examines the narrative/ data as a dyad and in exclusivity. Rather, we propose that there are additional factors or contexts in this relationship, such as code, disciplinary contribution and bias.

Presner’s work is perhaps the one that springs to mind most immediately in respect to the narrative/ data debate; not only because it is of relevance to the KPLEX project outline, but specifically in relation to his argument, noted previously, that datafication "has the effect of turning narrative into data":

> what goes missing in the “pursued objectivity” of the database is narrativity itself; from the dialogical employment of the events in sentences, phrases, and words in response to the interviewer's questions, to the tone, rhythm, and cadence of the voice, to the physical gestures, emotive qualities, and even the face itself.

Presner presents his examination of the relationship between narrative and data in “Ethics of the Algorithm” as

> an analysis of the *computational genre* itself in historical representation. This includes databases, structured data and queries, and all the algorithmic means in which data is mined, analyzed, and visualized. It seeks to locate the ethical in digital and computational modalities of representation.

Presner’s piece does do this, but in doing so it exposes other problems in terms of the suppositions made regarding what ethics are the “correct” ethics and why, and who gets to decide this. There is a need “to locate the ethical in digital and computational modalities of representation” but it is misdirected to suggest that these approaches themselves are somehow antithetical to ethics as it stands; these computational practices are already reflective of human subjectivities. Teasing out the full spectrum of relationship(s) that exist between data and narrative would greatly contribute to this area of research.

There is an awareness that context (both native and imposed) influences how we interpret, treat, and identify data. This has been discussed previously, but it is nonetheless necessary to acknowledge the modest critical awareness that is context and its influence on data. Borgman observes that data “exist in a context, taking on meaning from that context and from the perspective of the beholder. The degree to which those contexts and meanings can be represented influences the transferability of the data." Van Es and Schäfer classify datafication as the "the new empirical":

> Because datafication is taking place at the core of our culture and social organization, it is crucial that humanities scholars tackle questions about how this process affects our understanding and documentation of history, forms of social interaction and organization, political developments, and our understanding of democracy.
>
> The unquestionable allure of new forms of empiricism makes it important for us to continue to acknowledge that humanities scholars’ epistemological assumptions are different from those of their counterparts in the hard sciences.

Context was also cited as the de facto approach towards dealing with uncertainty and ambiguity in data. That this was such a prevalent response from researchers with different disciplinary backgrounds and with different research interests highlights the need for this approach to be regulated, standardised and further advocated and disseminated at an institutional level.

The datafication of the humanities:

> various attempts are being made to build ‘digital’ versions or extensions of long-established disciplines, this encounter marks a moment of destabilization and deterritorialization, a moment that implies significant contingency and different possible outcomes.

Having humanities researchers too readily adopted practices and protocols from the Sciences without thinking about the repercussions of these methodological approaches is problematic. This crossover between computer science practice, science, and the humanities appears to be one directional, with computational and scientific methods spilling

into the humanities without an equal but opposite response that sees humanities methodologies start to spread into/ influence computer science/ the sciences in return. This is a factor noted by Edmond in *Will Historians Ever Have Big Data.*

> If digital technology is set to change the way scholars work with their material, how they 'see' it and interact with it, a pressing question is how these methods affect the way we generate, present and legitimize knowledge in the humanities and social sciences. In what way are the technical properties of these tools constitutive of the knowledge generated? What are the technical and intellectual skills we need to master? What does it mean to be a scholar in a digital age? To a large extent, the answers to these questions depend on how well we are able to critically assess the methodological transformations we are currently witnessing.

There is a clear lack of understand of the computational turn among humanities researchers together with a reluctance to fully respond to it, acknowledge it, or embrace it. More detailed provenance, greater transparency regarding the transformations applied to data, and increased accountability on the part of those responsible for the curation of data would help counteract this distrust and dissuade insular behaviour among humanities researchers or researchers that lack fluency in computer science methodologies. Researcher accountability was highlighted not only as a key factor in assessing the validity or reliability of data, but as a key concern among researchers. The provision of adequate training and CPD regarding Best Practice apropos data curation, processing, and analysis are additional recommendations.

The impact of the one-directional crossover between the sciences and humanities in terms of methodologies, techniques and software from computer science (in particular) being incorporated into humanities research has a knock-on effect in the form of a comparable lack of uptake or adoption of humanities research methodologies among computer scientists, with Manovich asking "Why do computer scientists rarely work with large historical data sets of any kind?" In addition, however, there is a comparable one-directional crossover of methodologies between computer science and statistics: "Looking at the papers of computer scientists who are studying social media data sets, it is clear that their default approach is statistics." Where this becomes problematic is in the fact that, rather like humanities researchers appropriating and using computer science methodologies in ways that differ from those within the discipline of computer science, computer scientists themselves, as noted below, employ statistics different to those within other disciplines:

> Computer scientists studying social media use statistics differently than social scientists. The latter want to explain social, economic or political phenomena. Computer scientists are generally not concerned with explaining patterns in social media by referencing some external social, economic or technological factors.

It is likely then that should crossover between the humanities and computer sciences occur, the methodologies appropriated will likely be employed in atypical ways. This is why interdisciplinary collaboration and clear communications is so important.

It is impossible to an expert in the necessary cross-disciplinary fields, and therefore interdisciplinary collaboration must be encouraged and trust across disciplines fostered. This is a particular problem for big data and the humanities, particularly because what we are

being asked to become expert in is so divergent from our comfort zones. That said, as Edmond notes, there is a clear imbalance in terms of the number of humanities researchers being presented with the opportunity to acquire computer science skills versus the number of computer scientists presented with the opportunities to acquire skills from the humanities:

> On the other side, we should also be querying the imbalance in upskilling opportunities: there are many, many training programmes, summers schools, web resources and the like inviting humanists to learn programming skills, but where is the summer school introducing humanistic methods to computer scientists?

Taking on additional training to acquire at least a competency in or appreciation for cross-disciplinary practices would help to establish and foster trust regarding data practices across disciplines. But as both Lipworth et al and Rieder and Röhle note, even experts in these disciplines may struggle with the complexities of these digital frameworks:

> The complexity of analytic models and predictive algorithms may, for example, limit the capacity for the public, and even experts, to interpret and question research findings. This can cause real harm if people act on false predictions (including, but in no way limited to those stemming from 'incidental findings') [...] or lose sight of ethically important contextual nuances that are obscured by big data analyses.

> It is very naive to believe that anybody who has not had considerable training in both programming and simulation modelling can say anything meaningful about how ForceAtlas2 is implementing the force direction concept differently from its historical and conceptual ancestor, the work of Fruchterman and Reingold; and much less how these differences affect spatialisation in concrete circumstances. How will properties of nodes and topological structure affect positions on the map? Which aspects of the latent structures in the data does the diagram reveal?

The need to resist cherry picking data to adhere to pre-established research hypotheses or desired narratives, was also repeatedly stressed throughout the interviews. In this concluding discussion I would like to raise the question as to whether converting narrative to "data" be considered an act of cleaning; the implications of which are, as yet, epistemologically unexplored/ uncertain? How do we retain cognisance of that which has been scrubbed away? Also: what is being cleaned from the material in order to create data or narrative is—as Edmond notes in *Will Historians Ever Have Big Data* etc.—precisely the material humanities researcher thrives on.

> There are acknowledged widespread misconceptions around objectivity and data (collection, curation, cleaning). Again as Schäfer and van Es note, we have misconceptions about objectivity not only “In academic research, but also in many sectors of business and other areas of society at large, data analysis unfolds via computer interfaces that display results that users often mistakenly regard as objective assessments.”

Our current enthusiasm for computer-aided methods and data parallels the technology-induced crisis in representation and objectivity analysed by Daston and Galison. Their concerns must be taken into account in order to critically reflect upon the purported objectivity of computer-calculated results and visualizations.

> In the absence of readily legible clues as to their epistemic foundations, computational research tools are often assigned such values as reliability and transparency (Kitchin 2014: 130). As Rieder and Röhle observe, the automated processing of empirical data that they enable seems to suggest a neutral perspective on reality, unaffected by human subjectivity (2012: 72). Drucker, a specialist in the history of graphics, makes a similar point, focusing more closely on practices of data visualization. She argues that the tools used for this purpose are often treated as if the representations they render provide direct access to 'what is'. This way, the distinction between scientific observation ('the act of creating a statistical, empirical, or subjective account or image') and the phenomena observed is being collapsed (Drucker 2014: 125; see also Drucker 2012: 86).

There is also a need to radically reassess what exactly we want our databases to do. van Zundert, Antonijević, and Andrews stress the fact that "A practical examination and theoretical discussion of how software reflexively interacts with humanities research remains an urgent necessity." Do we want our digital research environments to mimic, or continue to mimic analogue methodology in the form of digital skeumorphisms that are themselves flawed and prone to error and human bias? Is it possible to design something that goes beyond bias? Many databases are organised and managed by systems that are "remarkably literalist" Is there an awareness among computer scientists of the implications of this literalism? There is a clear need to explore other forms of Knowledge Representation Schemes and alternatives to metadata, information management systems, organisational structures and search functions, forms that may be capable of overcoming the bias of their human engineers. Are we looking for a database that is truly post-human then? In that it is not only fundamentally without bias, but it evades the possibility of there being human interference, human hierarchising, or human bias? This imposition of no bias is in itself a bias of course. So rather than there being an ethics of the algorithm, wherein the ethics are determined by human, we would have an ethics imposing database, a database/ algorithm that performs ethical unbias, a de-biasing (de-humanising) machine.

6. Recommendations

Recommendation 1: Encouraging co-operation, mutual understanding and trust between researchers across disciplines, and between researchers and developers.

Recommendation 2: Encourage Researcher Accountability among researchers regarding what their data is, where it has come from, and what they have done to it.

Too many researchers are taking for granted what "data" are, and this is no longer sustainable as the European research landscape moves towards greater integration and inter-disciplinary connectivity. Data should be defined clearly in publications/ proceedings, on a case by case basis. Clarity of terms and clear communication is also required for surrounding weighted terms that are at risk of being misinterpreted. It was observed by a number of interviewees that consensus regarding data could perhaps be reached on a modular level, among smaller research groups or within a contained group.

Recommendation 3: Provide training for computer scientists or engineers involved in the development and provision of technology for researchers in disciplines outside of the computer sciences that may have different conceptions of what data are. Concordantly, provide training for researchers with no background in or understanding of the computer sciences

Recommendation 4: Develop a cross-discipline taxonomy for the curation and provenance of complex data.

Recommendation 5: The provision of training that foregrounds the ethics of responsible narrativisation/ narrative usage when it comes to the dissemination of information gleaned from data. There is a need for clarity regarding the function of narrative apropos the dissemination of data, of the role of narrative as a powerful (and easily manipulated) tool for explicating data. This is particularly important when it comes to complex data, sensitive data, or where the narrative is targeting non-specialist audiences who may not be capable of verifying the claims of the narrative by studying the data for themselves. The role of narrative in the explication of data needs to be further acknowledged among the computer science community. The prevalence of this confusion regarding the relationship between data and narrative points towards a need to clarify the functionality of narrative as it is employed as a tool by researchers to communicate their data, particularly as it is understood among researchers working within the computer sciences. The prevalence of the term datafication has been observed across all modules in this project. There was a notable absence of an equivalent term for the process of narrativising data, yet this phase is a key, if not integral, phase in every research project. The narrativization of data is a facet of the data lifecycle that, like data collection, curation, processing, and sharing, must be regulated, which would involve the provision of Best Practice guidelines to encourage the ethical usage of narrative in the dissemination of data.

7. Annex 1: Interview Questions.

Section One. Positioning Your Data Activities.

1. What is your field of study, and research interests? What is your object of your research?

2. Describe your research process: what are the inputs, what tools or processes do you use to interrogate them? How do you formulate the results of that interrogation into findings or insights, and what do they look like?

3. What theoretical underpinnings / basic assumptions guide your research/ your work/ your project/s?

Section Two. Assumptions and Definitions of Data.

4. How would you define data? Could this material, this data, be understood to mean different things in different situations?

5. In our research so far, we have read that data can be “pre-analytical, pre-factual,” “rhetorical,” “false,” that data “has no truth” it “resists analysis. [...] [it]cannot be questioned” it is "neither truth nor reality,” it is a “fiction,” an “illusion,” “performative,” “a sort of actor.” Data should be distinguished from “capta” meaning “taken” because data is something “given.” Data are “always preconstituted, shaped by the parameters for their selection” and that it possesses a capacity to reconfigure its environs and interlocutors.

6. Does this multiplicity surprise you? Do you find it problematic either theoretically or practically? Are there any specific instances you can recall?

Section Three. Data Cleaning, pre-processing and Processing:

7. You’ve described your research processes to us already in general terms, now we would like to focus on the transformations that either have already been applied to it, or are applied by you to your data in the course of your research. Can you describe this process, in terms of what happens before you get your data, what happens before you use your data, what happens while you are using your data, and finally, what happens after you have used your data?

8. Are these largely predictable processes, or do unexpected things happen? If so, can you give examples?

9. In these processes what (if anything) gets removed? What are the reasons for these materials being left out? Would you expect any other researcher to also choose to exclude these materials, irrespective of the project or programmer?

10. What happens to this excess or removed material? Is it stored anywhere? Are these changes reversible? Are they traceable/ trackable? Do you document them? Discuss them in your publications?

Section Four: Messy data/ difficult material:

11. Are you ever presented with data that is difficult to process? If so, what makes it difficult and how do you respond to those challenges?

We would like to present you now with three different scenarios of how data might be ‘difficult.’

12. Ambiguous data To be ambiguous means something can be interpreted in more than one way, that it can have more than one meaning.

Example:

The girl is crying.	She is crying because she is sad. She is crying because she is happy. She is crying because she is in pain.

Do you ever encounter data (material) that is ambiguous? If so, what makes it ambiguous? Can you give any examples? How do you deal with ambiguity?

13. Contradictory data: To be contradictory or to contradict means to say or suggest the opposite of what you have just said or suggested.

Example:

I love you.	I hate you.
The treatment is working.	The treatment is not working.

Do you ever encounter data (material) that is contradictory? Can you give any examples? How do you handle this material? Do you acknowledge it in your end results or publications? Do you remove it from your datasets?

14. Uncertain data. When something is uncertain its meaning is unresolved or unresolvable.

Example: Two people look at a circular shape drawn on a page: ○

It is a football, says one.	It is a moon, says the other.

Do you ever encounter data that can be similarly interpreted in different ways in your research? If so, can you give an example? How do you deal with or manage uncertainty or complexity when you encounter it in your research?

15. Do you think uncertainty or complexity in data is a good, bad or neutral thing? How do you respond to it when you recognize it?

Section Five: Relationship between data and narrative

16. How is your data structured/ organised? What factors influence how you set up your data structures? Can you give an example?

17. We'd like to ask you a final question about narrative.

A narrative is an account of connected events, or an account that argues for the connectedness of events. This account can be text-based (ie. an academic paper or conference presentation, a story) or it can be visual (ie. a graph or visual representation of data, a photograph).

Examples:

The patient ingested the drug and their condition improved. The drug helped to alleviate the condition.

Plants that are not watered will die. Plants need water.

Lots of people attended my wedding. We have photos of all the guests.

———

How does your data differ from a narrative? Do you consider there to be a relationship between data and narrative? If so, what kind of relationship?

Thank you very much for your responses!

7. Annex 2: WP2 Code List for the Qualitative Analysis of Interviews Used in Atlas.ti

Code Name	**Code Definition**	**Examples**
Acknowledging/ not acknowledging removal of data		
Ambiguous data	Data that can be interpreted in more than one way, data that can have more than one meaning	
Assumptions	Something that is accepted as true, without proof	Data definitions; DIKW hierarchy; Theoretical Underpinnings

Contradictory data	Data that is contradictory, opposite meanings within a dataset	Contradictory outliers
Co-op/ lack of co-op researchers & developers	Cooperation or lack of cooperation between researchers and developers in the development of tools	
Context dependency	Dependency of interpretations on the context of data	
Data/ algorithm	Relationship between data and algorithm(s).	
Data analysis	How data is analysed/ How researchers approach the analysis of data.	Computer Assisted Tools; Context dependency; Data manipulation; Disambiguation; Human-machine interaction; Inter-annotator agreement; Semantic Evaluation; Semantic Web technology (reasoner); Source of data; Tacit Knowledge; Tagging; Word sense dis-ambiguity
Data cleaning	Removing incorrect or inconvenient elements from the available data, supplying missing information and formatting it so that it fit with other data	Ignored; Leave the outliers out; Outlaw examples; Reformat; Removal of data; Split it apart; Strip out; Taken as noise; Thrown out; Washing.
Data collection	How data has been collected	
Data definitions	Definitions given for data	
Data/ information	Relationship between data and information	

Data interpreting	How data is interpreted	
Data lifecycle	The lifecycle of data	Data collection; Data deprecation; Data use; Downstream; Upstream; Repurposed data; Transformations of data
Data loss	Loss of data in the process of data collection, processing, or analysis	
Data (pre-) processing	Data pre-processing or data processing	Alignment; Back propagation; Formatting; Normalization; Parsing; Reengineering; Reformatting; Semantic Web Technology; Sentence alignment; Reasoner; Simplifying; Tokenize.
Data Privacy/ Ethics	Privacy concerns regarding data/ Ethics of data usage or acquisition	Anonymising data; Data Privacy; Ethical Issues; Personal data; Privacy Issues
Data representation	How data is represented	
Data/ Researcher	Specificities of the relationship between the researcher and “their” data	
Data sparsity	Sparsity of data in the process of data collection, processing, or analysis	
Data use	How/ where data is used	Primary uses; Secondary uses; Downstream; Upstream; Repurposed data
DIKW	Data Information Knowledge Wisdom	DIKW hierarchy; DIKW Comorbidity

Discipline's contribution	Contribution of a discipline to interdisciplinary research	Disciplinary background; Disciplinary differences; Domain knowledge/ lack of domain knowledge; Evaluation criteria; Tacit Knowledge; Theoretical Underpinnings
Emerging research disciplines/ questions	New research disciplines/ New research questions that have not been asked before	
Evaluation criteria	Criteria for the evaluation of data	Best practice; Tacit knowledge; Theoretical Underpinnings; Objectivity vs. Subjectivity
Interdisciplinary research		Co-op or lack of co-op researchers + developers; Humanities vs. Natural sciences (Differences between humanities and natural sciences (technically and data-driven approaches)); Knowledge gaps in the research field
Interpreting data	Approaches to the interpretation of data	Annotation; Context dependency; Data analysis; Data manipulation; Disambiguation; Inter-Annotator Agreement; Linked Open Vocabularies; Objectivity; Research aims; Sampling methods; Semantic Evaluation; Subjectivity; Tacit Knowledge; Tagging; Theoretical Underpinnings; Transformations of data; Word sense disambiguity
Data / Narrative	Relationship between data and narrative	DIKW

Research challenges	Key challenges in conducting research, i.e. Lack of a congruent definition of a concept	Cultural/language differences; Co-op or lack of co-op researchers + developers; Big Data; Comorbidity; Data loss/ data sparsity; DIKW hierarchy; Knowledge gaps in the research field; Lack of clarity re Best Practice; Not according to Best Practice; Relationship between algorithms and data; Research objects; Research policies; Sampling methods; Small Data; Source of data; Subjectivity; Tacit Knowledge; Validity
Researcher specifics	Specific skills/ background / tacit knowledge of the researcher	Cultural background; ; Cultural/language differences; Disciplinary background; Discipline's contribution; Domain knowledge/ lack of domain knowledge; Disciplinary differences; Emerging research disciplines; Emerging research questions; Evaluation criteria; Humanities vs. Natural sciences; Mediation; Tacit Knowledge;
Research Processes	Methods applied to conduct research	Algorithms; Data collection; Data cleaning; Data manipulation; Experiments; Interviews; Machine Learning; Observations; Sampling methods; Sensors; Surveys; Theoretical Underpinnings; Machine translation; Natural Language Processing
Objectivity vs. Subjectivity	Issues relating to objectivity and subjectivity in relation to research data.	Cultural background; Cultural/language differences

Source of data	Where the data originated from	Social media: Facebook/ Twitter; Media analysis; User Generated Content; Text analysis
Terminological comorbidity	Comorbidity of terms/ meanings; Interchangeability or comorbidity of the terms data/ information, etc. Or, for example, facets of the tags used to tag data referred to as “information.”	Data definitions; data/ information; DIKW comorbidity
Transformations of data	Transformations applied to research data	
Types of data	Types of Data referred to by the researcher.	Ambiguous data; anonymous data; bad data; big data; bilingual data
Uncertain data	Meaning of data is unresolved	
Validity/ Reliability	Accuracy of measurement; Consistency of measurement, i.e. inter-coder agreement.	

CODE GROUPS	**Code Group Definition**	**References**	**Codes within the Group**	**EXAMPLES**
ASSUMPTIONS ABOUT DATA	Assumptions made about data		Assumptions; Comorbidity; Context dependency; DIKW hierarchy; Validity.	

CHALLENGES			Co-op or lack of co-op researchers + developers; Data Privacy/ Ethics; Emerging research disciplines/ questions; Research challenges.	
DATA RELATIONSHIPS	Relationships between data and other actors		Comorbidity; Data/ algorithm; Data/ information; Data/ narrative; Data / researcher	
DATA MANIPULATIONS/ TRANSFORMATIONS	Transformations made to data throughout its lifecycle		Data analysis; Data collection; Data cleaning; Data interpreting; Data loss/ data sparsity; Data (pre-) processing; Data representation; Data use; Source of data.	Algorithm; Annotation; Artificial Intelligence; Classification; Encoding scheme; Inter-annotator agreement; Linked Open Vocabularies; Removal of data; Semantic Web Technology; Tagging
DIFFICULT DATA	Data that is particularly difficult to work with		Ambiguous data; Contradictory data; Uncertain data.	Bad data; Big Data; Bilingual data; Chaotic data; Contradictory outliers; Dark data; Hidden data; Missing data; Noisy data; Outdated data.

7. Annex 3: WP2 Data Mining Results

7.3.1.i) Journal of Big Data 2014-2017, merged table of results for iterations of the term “data.”

Term: “data”	***Count***	***Length***	***Trend***
data 4 we finally test whether big data capabilities are of crucial importance for the financial returns linked to big data projects we find indeed that mastering capabilities at scale are necessary to generate returns above cost of capital for big data in the telecom industry	*2*	*46*	*2*
data investments 3 third using a joint model of big data adoption and of returns on adoption we try to explain key drivers of this variance in big data performance consistent with the theory of technology adoption e.g	*2*	*38*	*2*
data contribution to total telecom profit is minor but in line with its relative size of investment in total telecom spent and generates a productivity impact aligned with other research tambe	*2*	*31*	*2*
data ingestion with and without data transformation performance results of intermediate data transformation using a mapreduce job and performance results of a simple analytic computation with and without data transformation sections	*2*	*31*	*2*
data we have designed and developed two contention avoidance storage solutions collectively known as bid bulk i o dispatch in the linux block layer specifically to suit	*2*	*27*	*2*
data from the 2010 census demographic and economic surveys bureau of labor statistics and center for medicare services	*2*	*18*	*2*
data https www.cms.gov research statistics data andsystems statistics trends and reports medicare provider charge data downloads	*2*	*16*	*2*
data pipeline in the context of the requirements identified in rq1 of this study part	*2*	*15*	*2*
data analytics bda is able to deliver predictions based on executing a sequence of	*2*	*14*	*2*

data there were no contextual anomalies cleared additionally when removing the day and timeofday	*2*	*14*	*2*
data is managed between heterogeneous tiers of storage in enterprise data center environment	*2*	*13*	*2*
data source dropdown select 2010 census in the data variable dropdown select	*2*	*12*	*2*
data access the i o stack still remains volatile the	*2*	*10*	*2*
data application logs clickstream logs whether data emails contracts geographic	*2*	*10*	*2*
data upload and graph generation times based on number of	*2*	*10*	*2*
data analytics the adoption of big data analytics is	*2*	*9*	*2*
data and task parallelism when processing an ou pipeline	*2*	*9*	*2*
data centers would also lead to energy footprint reduction	*2*	*9*	*2*
data definitions are expressed in bio medical scientific publications	*2*	*9*	*2*
data points that are considered abnormal when viewed against	*2*	*9*	*2*
data projects further as found in theory of production	*2*	*9*	*2*
data intensive bioinformatics workflows in hybrid cloud environments	*2*	*8*	*2*
data the content detector was able to find	*2*	*8*	*2*
data collection instructions from the site manager	*2*	*7*	*2*
data refers to the infrastructure and technologies	*2*	*7*	*2*
data scalability of models and distributed computing	*2*	*7*	*2*
data access part of the simulation	*2*	*6*	*2*

data acquisition preprocessing analysis and interpretation	*2*	*6*	*2*
data centers experiencing big data workloads	*2*	*6*	*2*
data dependencies of a number of	*2*	*6*	*2*
data processing components in the pipeline	*2*	*6*	*2*
data processing frameworks such as hadoop	*2*	*6*	*2*
data projects source bughin 7 sample	*2*	*6*	*2*
data reports spreadsheets and social media	*2*	*6*	*2*
data representations from large volumes of	*2*	*6*	*2*
data summarization graphical representation dimension reduction	*2*	*6*	*2*
data tagging fast information retrieval and	*2*	*6*	*2*
data 3 5 8 10	*2*	*5*	*2*
data analytics and 2 how	*2*	*5*	*2*
data analytics in addition to	*2*	*5*	*2*
data analytics including learning from	*2*	*5*	*2*
data analytics tamuse thus supporting	*2*	*5*	*2*
data big data analytics is	*2*	*5*	*2*
data collection storage and analytics	*2*	*5*	*2*
data deep learning algorithms are	*2*	*5*	*2*
data driven industrial analytics applications	*2*	*5*	*2*

data exists within the same	*2*	*5*	*2*
data from across the factory	*3*	*5*	*3*
data from the factory to	*2*	*5*	*2*
data is stored in the	*2*	*5*	*2*
data management get amplified in	*2*	*5*	*2*
data mining and machine learning	*6*	*5*	*6*
data pipeline presented in this	*2*	*5*	*2*
data sets that did not	*2*	*5*	*2*
data to be processed by	*2*	*5*	*2*
data to train a classifier	*2*	*5*	*2*
data transformation occurs within the	*3*	*5*	*3*
data with respect to the	*2*	*5*	*2*
data abstractions and representations	*3*	*4*	*3*
data analytics as a	*2*	*4*	*2*
data analytics presents a	*2*	*4*	*2*
data analytics presents an	*2*	*4*	*2*
data and concept drift	*3*	*4*	*3*
data and materials the	*2*	*4*	*2*
data and parameter selection	*2*	*4*	*2*

data and smart manufacturing	*2*	*4*	*2*
data archive and the	*2*	*4*	*2*
data as shown in	*2*	*4*	*2*
data as well as	*2*	*4*	*2*
data because they are	*2*	*4*	*2*
data big data is	*3*	*4*	*3*
data characteristics of the	*2*	*4*	*2*
data contribution to profit	*4*	*4*	*4*
data dealing with high	*2*	*4*	*2*
data driven analytics applications	*2*	*4*	*2*
data driven companies e.g	*2*	*4*	*2*
data driven systematic approach	*2*	*4*	*2*
data elements which may	*2*	*4*	*2*
data for both measurements	*2*	*4*	*2*
data from different sources	*2*	*4*	*2*
data from which the	*2*	*4*	*2*
data has already been	*2*	*4*	*2*
data if it is	*2*	*4*	*2*
data in addition to	*2*	*4*	*2*

data in industrial environments	*3*	*4*	*3*
data in order to	*3*	*4*	*3*
data in other words	*2*	*4*	*2*
data in real time	*4*	*4*	*4*
data integration and contextualisation	*3*	*4*	*3*
data is characterized by	*2*	*4*	*2*
data is created every	*2*	*4*	*2*
data is to be	*3*	*4*	*3*
data location filtering layer	*2*	*4*	*2*
data mining and statistical	*2*	*4*	*2*
data pipeline apache flume	*3*	*4*	*3*
data pipeline focuses on	*2*	*4*	*2*
data pipeline so the	*2*	*4*	*2*
data points for the	*2*	*4*	*2*
data points in the	*2*	*4*	*2*
data points which are	*2*	*4*	*2*
data problems such as	*2*	*4*	*2*
data representations in a	*5*	*4*	*5*
data representations obtained from	*2*	*4*	*2*

data shared resource environments	*2*	*4*	*2*
data source adapted from	*2*	*4*	*2*
data source dropdown for	*2*	*4*	*2*
data stored in the	*3*	*4*	*3*
data stream clustering algorithms	*2*	*4*	*2*
data stream clustering in	*2*	*4*	*2*
data stream clustering method	*2*	*4*	*2*
data stream clustering solutions	*2*	*4*	*2*
data stream clustering which	*3*	*4*	*3*
data stream clustering with	*2*	*4*	*2*
data stream in the	*2*	*4*	*2*
data such as text	*2*	*4*	*2*
data ti or big	*2*	*4*	*2*
data to the cloud	*2*	*4*	*2*
data validity and reliability	*3*	*4*	*3*
data variable dropdown for	*2*	*4*	*2*
data which can be	*2*	*4*	*2*
data with deep learning	*2*	*4*	*2*
data 1 the	*2*	*3*	*2*

data 2014 1	*2*	*3*	*2*
data 2015 2	*5*	*3*	*5*
data 2016 3	*5*	*3*	*5*
data 2017 4	*5*	*3*	*5*
data ab or	*2*	*3*	*2*
data access for	*2*	*3*	*2*
data access is	*2*	*3*	*2*
data access requirements	*2*	*3*	*2*
data access time	*2*	*3*	*2*
data according to	*2*	*3*	*2*
data adoption in	*3*	*3*	*3*
data analysis and	*7*	*3*	*7*
data analysis problems	*2*	*3*	*2*
data analysis to	*2*	*3*	*2*
data analysts to	*2*	*3*	*2*
data analytics by	*2*	*3*	*2*
data analytics has	*2*	*3*	*2*
data analytics however	*3*	*3*	*3*
data analytics hypothesis	*4*	*3*	*4*

data analytics research	*4*	*3*	*4*
data analytics where	*2*	*3*	*2*
data analytics which	*3*	*3*	*3*
data analytics with	*2*	*3*	*2*
data and or	*2*	*3*	*2*
data and the	*3*	*3*	*3*
data and then	*3*	*3*	*3*
data application domains	*2*	*3*	*2*
data approaches in	*2*	*3*	*2*
data are being	*2*	*3*	*2*
data are of	*2*	*3*	*2*
data are stored	*3*	*3*	*3*
data as a	*5*	*3*	*5*
data at the	*2*	*3*	*2*
data aware optimization	*3*	*3*	*3*
data being stored	*2*	*3*	*2*
data between the	*2*	*3*	*2*
data by applying	*2*	*3*	*2*
data by which	*2*	*3*	*2*

data can also	*2*	*3*	*2*
data can be	*16*	*3*	*16*
data center applications	*2*	*3*	*2*
data center storage	*4*	*3*	*4*
data centers in	*2*	*3*	*2*
data cloud deployments	*2*	*3*	*2*
data collec on	*2*	*3*	*2*
data collection and	*3*	*3*	*3*
data collection is	*3*	*3*	*3*
data datasets from	*2*	*3*	*2*
data definitions and	*3*	*3*	*3*
data definitions in	*2*	*3*	*2*
data deployments as	*2*	*3*	*2*
data deployments section	*4*	*3*	*4*
data deployments sections	*2*	*3*	*2*
data distribution a	*3*	*3*	*3*
data does not	*3*	*3*	*3*
data elements and	*3*	*3*	*3*

data field in	*2*	*3*	*2*
data for a	*2*	*3*	*2*
data for analytics	*5*	*3*	*5*
data for example	*6*	*3*	*6*
data for the	*8*	*3*	*8*
data formats and	*2*	*3*	*2*
data freshness and	*2*	*3*	*2*
data freshness by	*2*	*3*	*2*
data freshness value	*2*	*3*	*2*
data from a	*3*	*3*	*3*
data from archives	*2*	*3*	*2*
data from sources	*2*	*3*	*2*
data has been	*8*	*3*	*8*
data have been	*2*	*3*	*2*
data import utility	*2*	*3*	*2*
data in 2014	*2*	*3*	*2*
data in a	*6*	*3*	*6*
data in each	*2*	*3*	*2*
data in stream	*2*	*3*	*2*

data in telecom	*2*	*3*	*2*
data in these	*3*	*3*	*3*
data in this	*4*	*3*	*4*
data incremental rate	*2*	*3*	*2*
data ingestion from	*2*	*3*	*2*
data instead of	*3*	*3*	*3*
data intensive part	*4*	*3*	*4*
data into a	*3*	*3*	*3*
data into hdfs	*2*	*3*	*2*
data into the	*3*	*3*	*3*
data is a	*2*	*3*	*2*
data is accessed	*3*	*3*	*3*
data is available	*4*	*3*	*4*
data is collected	*2*	*3*	*2*
data is generated	*2*	*3*	*2*
data is its	*2*	*3*	*2*
data is located	*2*	*3*	*2*
data is modeled	*2*	*3*	*2*
data is not	*6*	*3*	*6*

data is still	*2*	*3*	*2*
data is the	*6*	*3*	*6*
data is then	*2*	*3*	*2*
data is transformed	*2*	*3*	*2*
data is typically	*2*	*3*	*2*
data is very	*2*	*3*	*2*
data it has	*2*	*3*	*2*
data it is	*6*	*3*	*6*
data loaded during	*2*	*3*	*2*
data management and	*5*	*3*	*5*
data management methodology	*2*	*3*	*2*
data management paradigm	*2*	*3*	*2*
data may be	*2*	*3*	*2*
data mining community	*2*	*3*	*2*
data mining data	*2*	*3*	*2*
data monetization is	*2*	*3*	*2*
data not shown	*4*	*3*	*4*
data obtained from	*2*	*3*	*2*
data of the	*2*	*3*	*2*

data on profit	*2*	*3*	*2*
data on the	*3*	*3*	*3*
data or datasets	*2*	*3*	*2*
data organization methodology	*2*	*3*	*2*
data over time	*3*	*3*	*3*
data per day	*2*	*3*	*2*
data pipeline and	*2*	*3*	*2*
data pipeline architecture	*1 1*	*3*	*1 1*
data pipeline but	*2*	*3*	*2*
data pipeline for	*6*	*3*	*6*
data pipeline is	*3*	*3*	*3*
data pipeline this	*2*	*3*	*2*
data pipeline to	*4*	*3*	*4*
data pipeline was	*2*	*3*	*2*
data pipeline with	*3*	*3*	*3*
data point is	*2*	*3*	*2*
data points into	*2*	*3*	*2*
data points where	*2*	*3*	*2*

data pre processing	*3*	*3*	*3*
data processing and	*2*	*3*	*2*
data processing component	*2*	*3*	*2*
data projects are	*3*	*3*	*3*
data quality and	*2*	*3*	*2*
data quality as	*2*	*3*	*2*
data quality is	*2*	*3*	*2*
data rather than	*3*	*3*	*3*
data representations and	*4*	*3*	*4*
data representations for	*4*	*3*	*4*
data represented by	*2*	*3*	*2*
data returns exhibit	*2*	*3*	*2*
data science and	*2*	*3*	*2*
data set as	*2*	*3*	*2*
data set is	*2*	*3*	*2*
data set of	*2*	*3*	*2*
data sets and	*2*	*3*	*2*
data sets are	*2*	*3*	*2*
data sets be	*2*	*3*	*2*

data sets being	*2*	*3*	*2*
data sets to	*2*	*3*	*2*
data should be	*3*	*3*	*3*
data simulation and	*2*	*3*	*2*
data software to	*2*	*3*	*2*
data source e.g	*2*	*3*	*2*
data source for	*2*	*3*	*2*
data sources data	*2*	*3*	*2*
data sources to	*2*	*3*	*2*
data sources used	*2*	*3*	*2*
data specifically the	*2*	*3*	*2*
data stream as	*2*	*3*	*2*
data stream is	*2*	*3*	*2*
data stream mining	*3*	*3*	*3*
data stream we	*2*	*3*	*2*
data streams the	*2*	*3*	*2*
data tagging our	*2*	*3*	*2*
data technologies and	*2*	*3*	*2*
data technologies are	*2*	*3*	*2*

data technologies hypothesis	*4*	*3*	*4*
data technologies in	*2*	*3*	*2*
data than the	*2*	*3*	*2*
data that are	*3*	*3*	*3*
data that have	*2*	*3*	*2*
data that is	*1 1*	*3*	*1 1*
data that were	*2*	*3*	*2*
data the data	*2*	*3*	*2*
data the main	*2*	*3*	*2*

7.3.1.ii) Journal of Big Data 2014-2017, merged table of results for iterations of the term “big.”

Term: “big”	Count	Length	Trend
big data 4 we finally test whether big data capabilities are of crucial importance for the financial returns linked to big data projects we find indeed that mastering capabilities at scale are necessary to generate returns above cost of capital for big data in the telecom industry	2	47	2
big data investments 3 third using a joint model of big data adoption and of returns on adoption we try to explain key drivers of this variance in big data performance consistent with the theory of technology adoption e.g	2	39	2
big data contribution to total telecom profit is minor but in line with its relative size of investment in total telecom spent and generates a productivity impact aligned with other research tambe	2	32	2
big data we have designed and developed two contention avoidance storage solutions collectively known as bid bulk i o dispatch in the linux block layer specifically to suit	2	28	2
big data analytics bda is able to deliver predictions based on executing a sequence of	2	15	2
big data analytics the adoption of big data analytics is	2	10	2
big data definitions are expressed in bio medical scientific publications	2	10	2
big data projects further as found in theory of production	2	10	2
big data refers to the infrastructure and technologies	2	8	2
big data projects source bughin 7 sample	2	7	2
big data analytics and 2 how	2	6	2

big data analytics in addition to	2	6	2
big data analytics including learning from	2	6	2
big data analytics tamuse thus supporting	2	6	2
big data analytics presents a	2	5	2
big data analytics presents an	2	5	2
big data and smart manufacturing	2	5	2
big data contribution to profit	4	5	4
big data is characterized by	2	5	2
big data pipeline focuses on	2	5	2
big data problems such as	2	5	2
big data shared resource environments	2	5	2
big data source adapted from	2	5	2
big data 1 the	2	4	2
big data 2014 1	2	4	2
big data 2015 2	5	4	5
big data 2016 3	5	4	5
big data 2017 4	5	4	5
big data adoption in	3	4	3
big data analytics as	5	4	5

big data analytics by	2	4	2
big data analytics has	2	4	2
big data analytics however	3	4	3
big data analytics hypothesis	4	4	4
big data analytics research	4	4	4
big data analytics where	2	4	2
big data analytics which	3	4	3
big data analytics with	2	4	2
big data application domains	2	4	2
big data approaches in	2	4	2
big data as a	3	4	3
big data big data	2	4	2
big data can be	2	4	2
big data cloud deployments	2	4	2
big data definitions and	3	4	3
big data definitions in	2	4	2
big data deployments as	2	4	2
big data deployments section	4	4	4
big data deployments sections	2	4	2

big data field in	2	4	2
big data for analytics	3	4	3
big data has been	2	4	2
big data in 2014	2	4	2
big data in telecom	2	4	2
big data is not	2	4	2
big data is the	4	4	4
big data on profit	2	4	2
big data pipeline architecture	8	4	8
big data pipeline for	5	4	5
big data projects are	3	4	3
big data returns exhibit	2	4	2
big data should be	2	4	2
big data software to	2	4	2
big data technologies and	2	4	2
big data technologies are	2	4	2
big data technologies hypothesis	4	4	4
big data technologies in	2	4	2
big data themes in	2	4	2

big data use case	5	4	5
big data use cases	9	4	9
big diversity in data	2	4	2
big data 15	2	3	2
big data 3	2	3	2
big data acquisition	2	3	2
big data an	2	3	2
big data analysis	3	3	3
big data another	2	3	2
big data applications	9	3	9
big data approach	2	3	2
big data architecture	3	3	3
big data are	6	3	6
big data benchmarks	2	3	2
big data but	2	3	2
big data by	3	3	3
big data characteristics	2	3	2
big data contributes	2	3	2
big data corpus	3	3	3

big data data	3	3	3
big data definition	2	3	2
big data domains	6	3	6
big data environment	2	3	2
big data environments	6	3	6
big data have	2	3	2
big data infrastructure	3	3	3
big data introduction	2	3	2
big data it	6	3	6
big data methodology	2	3	2
big data or	2	3	2
big data org	2	3	2
big data processing	2	3	2
big data quality	2	3	2
big data requirements	2	3	2
big data requires	2	3	2
big data research	3	3	3
big data researchers	2	3	2
big data science	3	3	3

big data specific	2	3	2
big data systems	5	3	5
big data talent	2	3	2
big data talents	2	3	2
big data technology	4	3	4
big data that	3	3	3
big data the	12	3	12
big data this	3	3	3
big data thus	2	3	2
big data to	6	3	6
big data tools	2	3	2
big data volume	2	3	2
big data v’s	3	3	3
big data was	2	3	2
big data where	2	3	2
big data which	2	3	2
big data while	4	3	4
big data with	2	3	2

big data workloads	2	3	2
big enough to	2	3	2
big sensor data	5	3	5
big and	2	2	2
big datasets	5	2	5

7.4 Annex 4: A KPLEX Primer for the Digital Humanities

Definitions of terms relevant for the study of Big Data

The big data research field is full of terminology, some of which will be familiar to humanists, some of which may not. When it is familiar, it may also mean something completely different from what we might expect. What follows is an interpretive glossary, intended not so much to add one more claim to an authoritative voice so much as to translate some of these terms into a humanist's range of experience and frame of reference, as a reflection of the experiences of the KPLEX project. It is based on our experience of working across disciplines, not only humanities and computer science, but also equally distinct fields like Science and Technology Studies (STS). We share it at the encouragement of our project reviewers.

Please note also that in spite of our respect for the Latin roots of the word and its linguistically plural status, we use 'data' grammatically as a singular collective noun.

Algorithm – this is a set of rules according to which computational systems operate. Given the statistical underpinning of many big data algorithms, they can be sources of objective processing or introduce bias, depending on the model of the world on which they are based.

Artificial Intelligence (AI) – where computation-based systems are able to move beyond simple logical or rule-based results, or programmed using a training methodology, the system is said to demonstrate artificial intelligence.

Big Data – there is no single definition or measure of size that determines what data is big. Indeed, the most common definitions rely on a number of measures (often four, but many lists give more), not all of which actually measure size, but many of which are described by terms that begin with the letter 'V': eg. volume, variety, velocity, and veracity.

Capta – this term (generally credited to Johanna Drucker) exists as an alternative to 'data,' stemming from the Latin word for 'to take' rather than 'to give.' Use of it reflects the recognition that data is never 'given' but in fact always 'taken' by human agency.

Complexity - reflecting phenomena in a digital environment requires it to be expressed somehow in binary code of 1s and 0s. But we do not know how to capture and express all aspects of an object or phenomena in this way. Humans receive information in multimodal ways - through sight, smell, touch, hearing, for example. Much of the information we receive is understood only tacitly. All of these layers, tacit and explicit, mean that objects and phenomena in the real world generally possess a greater complexity than their digital surrogates (though digital systems themselves can possess their own complexity). In digitisation, we generally capture only part of a phenomenon, thereby reducing its complexity for the next user of that surrogate. How this choice of what to capture is made is also referred to as *tuning* the *signal to noise ratio* (with signal being what you think is important, and noise the unimportant or distracting elements).

Needless to say, your tuning will generally depend on your needs or values regarding a dataset.

Computational - a term that would be associated with *quantitative* (or mixed) research approaches, but referring not to research processes at the macro level, but to specific mathematical processing of numerical information as a component of such an approach where often an algorithm could be applied.

Context - the process of datification is by necessity a simplification. One aspect of this simplification is often the removal or reduction of context, that is the rich set of interconnections and relationships that any given object of phenomena possesses.

Curation - a process by which a selection or organisation of items is made from a larger possible collection. Can also refer to the presentation, integration or annotation of such a dataset.

Data - you need only read a bit of this report to realise that many communities define data in different ways: even internal consistency in these communities (or indeed in individual research papers) in the definition and use of this key term is often not the norm. It is therefore important in any interdisciplinary collaboration involving talk of data to ensure the parties involved make explicit what they mean by data, and what the implications of this working definition are.

Data Cleaning, also known as *data scrubbing* - a process by which elements are removed from a dataset or stream, generally because they foul the desired processing. This process is viewed by some communities as a central part of good research practice; others, however, view data cleaning as a form of *data manipulation* that erodes the credibility of research based upon it.

Datafication - a process generally understood as the rendering of original state objects in digital, quantified or otherwise more structured streams of information.

Data Manipulation - see *Data Cleaning*.

Data Scrubbing - see *Data Cleaning*.

Digital - this may seem obvious, but something is digital having been converted into a representation consisting of ‘1’s and ‘0’s (binary code) that can be read by a computer. Digital surrogates of analogue objects do not necessarily carry all of the properties of the original, however: for example, a page of text can be digitised as a photograph, but that doesn’t mean that the computer can read what is printed on the page, only that it has captured a sequence of lighter and darker pixels.

Documentation - this term has a rich history, but its formal use is generally attributed to Suzanne Briet. It is a process similar to *datafication*, but coined in an age before computers, referring to a number of technological and human processes that create a durable record of an object (for example, a geological formation, which cannot itself be

held in a museum), an event (such as a performance), or other human activity or natural phenomenon.

DIKW - a commonly used acronym to capture the relationships between different kinds or states of the building blocks of understanding. It stands for 'Data, Information, Knowledge, Wisdom.' KPLEX findings have shown that it is flawed in its lack of recognition of the positionality of knowledge states.

Information architecture - data must be organised according to some kind of framework, and the major and minor elements in this framework can be referred to as its architecture (just as a house will have a certain collection of rooms and hallways into which furniture - data - can be placed).

Machine learning – a process by which computational intelligence is developed by presenting a system with a large body of information, from which it extracts rules and patterns without direct human oversight of exactly what patterns it extracts. Also referred to as training. A common component of building AI.

Metadata - literally, data about data. Like catalogue information is to a collection of books, or a jukebox listing is to a collection of songs, it is a shorthand, often standardised, description of objects in a collection.

Narrative - this terms seems to have an interesting relationship to data, reflecting some of the common (erroneous) claims that data would be closer to objectivity than other forms by which information is communicated. Narrative is seen by many as the other side of that coin, the 'story' that someone either wants to or finds a way to tell using data (with a common implication that this 'story' does not reflect an objective position).

Native Data - this term can have a couple of meanings. In the context of commercial or other software development, it may refer to data formats that are most compatible with a particular system, or indeed created by it. For example, a Word document is most easily read in that software environment: viewing it through another word processor may cause errors or artefacts to appear. Native data is also, however, used to refer generally to data that is close to its original context of datification, as it exists/existed in its indigenous environment, the source context or environ where it was extracted from; so [original] data plus its environmental or source context.

Neural Networks - a form of computational system commonly associated with artificial intelligence.

Paradata - sometimes called 'data byproducts,' paradata are some of the most interesting forms of data (to the humanist, at least) because of their dependence and interrelatedness with other data streams. Notes in the margin of a book are a good example of paradata, as they document the connection between the thoughts of the reader and the text being read.

Processing - taking data and changing it, by filtering, reorganising, interrogating or otherwise applying some set of transformations (permanent or temporary) to it.

Provenance - a particular perspective on context, central in the world of museums, libraries and archives. Provenance refers to the record of where an object, collection or dataset has come from, and the places and 'experiences' (additions, transformations, losses, audiences, etc.) it has had since its original documentation.

Raw Data - what this phrase refers to is, as Sandra Gitelman puts it, an oxymoron. And yet the term persists. In theory, it refers to data that has not undergone any processing, or more specifically, any *further* processing, as the process of datafication in and of itself is a transformative process. It can also refer to data that has yet to be in any way interpreted or marked by human perspectives.

Source data - indicates the source of the data for any given research project or use. For some researchers, their source data may not be native, or raw; it may already be data proper and have undergone extensive processing, whether or not they recognise this is a part of how they situate their use of the data and their results.

Social Costs - since the time of Plato, it has been recognised that knowledge technologies bring both costs and benefits to their users. Writing may have led to the preservation and dissemination of many texts, but it also certainly weakened the habits and capabilities associated with the oral tradition. Digital technologies, such as data-driven research and big data collection have similar 'pharmakon'-like properties, meaning they can both help and hurt. In the contemporary context, however, the weighing of the 'costs' can be a challenging process, in particular because companies can generate profit by exploiting or at least skirting the edge of practices and products that may have a negative effects on individuals' health or overall social cohesion. It is therefore more important than ever that the impact of issues like identity, culture, privacy and cohesion are considered and (where necessary) defended against.

Standards - many communities agree to use the same sets of descriptors for certain features of a collection of objects or data. These tend to be agreed by communities, for example library standards such as Marc 21 or the standards for the sizes of sheets of paper based upon A5, A4, A3 etc. Community bodies can sometimes manage and validate standards (as the Society of American Archivists does for the EAD, or Encoded Archival Description); many others are described and promoted by the international standards agency ISO. There is overlap between standards in this sense and other tools for restricting and aligning how data is described, such as *controlled vocabularies* (one well-known example of this is the Getty terminology set, which provides standardised definitions of terms related to art and architecture).

Structured Data - in a database, small information elements will be organised and grouped at a granular level, like with like, such as a list of titles or names. *Unstructured data* will be more like narrative text, with no external order imposed to enhance processability.

Trust - this is a fundamental component of data driven research, big or otherwise, and means very different things to different people. Trust is a component of how we interact with digital environments and sources: it is the first thing any digital environment needs

to inspire in us (which usually occurs via proxies and heuristics, which may or may not actually speak to the basis on which a source should be trusted). A trusted environment or dataset may be seen to have *authority*, that is a claim to greater knowledge or experience than other possible sources.