iCLIC Data Mining & Data Sharing workshop: The Present and Future of Data Mining and Data Sharing in the EU

Southampton, UK, 23 September 2016

Robert Thorburn, Sophie Stalla-Bourdillon, Eleonora Rosati

Held at Southampton University's Highfield campus and hosted by iCLIC, an interdisciplinary core on Law, the Internet and Culture, the Data Mining and Data Sharing workshop brought together attendees and speakers from industry, government, academia and a range of disciplines alike. The workshop comprised of two sessions, each with a keynote and an associated panel. The first session was chaired by Eleonora Rosati and dealt with copyright and database rights, data mining and data sharing. The second session, chaired by Sophie Stalla-Bourdillon, focused on data protection, data mining and data sharing. The following report covers both sessions, associated panel discussions and the subsequent question and answer sessions.

First keynote: Julia Reda, Member of the European Parliament

With reference to the new European Commission proposals, MEP Julia Reda's presentation focused on new developments excluding issues of data protection, as this was to be dealt with extensively in the second session. Central to Reda's presentation was the question as to why a text and data mining (TDM) exception is needed at all when dealing with issues of copyright, as copyright should be concerned with the protection of creative works, whilst TDM deals with the extraction of facts and not the replication of intact works.

Starting with the Commission's initial goals for copyright reform, Reda noted the emphasis on making rules clearer, enabling Europe's digital potential by removing national silos and the positive impact thereof on research. Unfortunately, on the count of making the rules clearer the new Commission proposals seem to fall short of the mark, specifically on the issues of a new neighbouring right for publishers and a complication of the intermediary liability regime. On research and education though, the latest package performs better and includes new exceptions in these fields. The proposal includes three exceptions that are both mandatory and applicable EU-wide with a cross border effect. These are the illustration for teaching, but only in the digital environment; the preservation of works by libraries and archives (needed for the mass digitisation of collections); and lastly the TDM exception.

Focussing then on the TDM exception, Reda noted that there is a purpose-based limitation which applies the exception only to research, and has as sole beneficiaries research organisations. Research organisations are then also further defined as being organisations that perform the research themselves and are non-profits, or reinvest all profits back into the conduct of research; there has to be a public interest mission. By extension then, public private partnerships should be covered, the beneficiary requirement does however still present a significant hurdle. This is due to the beneficiary requirement being applied to all exceptions and as such it is not "*research in general, it is research by a research organisation, it is not TDM in general, it is TDM by a research organisation*". This presents a significant issue for the TDM exception regime in the UK which is limited to non-commercial purposes but not limited in terms of the beneficiary, meaning that individuals would also qualify for this exception. Also complicating the issues is that the Commission has chosen not to replace existing optional exceptions, but has rather added new mandatory exceptions. Therefore, national exceptions currently allowed will remain, but from a legal clarity point of view this situation is unsatisfactory.

Referring to the "The Council Conclusions on Open Science", Reda then pointed out several encouraging developments for TDM, including the Council's recognition of the importance of TDM, the need to facilitate TDM for all interested parties including citizen scientists and businesses (including SME's) to mine the results of publicly funded research they already have access to. This more progressive approach from the Council on copyright reform may then indicate future room for improvement in the legislative process. It was further noted though, that this proposed exception applied to reproduction rights only. As such, one could perform TDM and publish the results thereof, but not share the raw data with other scientists. This then makes it more difficult to conduct replication studies.

Returning to the proposed new neighbouring right for publishers, Reda noted that an additional hurdle for science may be present here. This hurdle not only extends to any publication with at least some journalistic content, but is also to be applied retrospectively to all publications from the past twenty years. Therefore, even completed research could be brought under this regime. Furthermore, this new neighbouring right extends to non-copyright materials, thereby making it even more onerous.

In sum, Reda noted that it would be difficult to make a blanket statement on the new proposals being either positive or negative. This is due to progress being made in some areas whilst significant resistance to change remains in others. This is in part due to the fact that activities

such as TDM should never have become copyright issues in the first place. These activities were perfectly legal and acceptable in the analogue environment, but copyright legislation has not kept pace with technological advancement. A case in point is the mandatory exception for reproductions that are technological and transient in nature, such as the technology employed in hearing aids. A seemingly logical exception which was in fact vehemently opposed at the time of introduction, especially by the publishing industry. Similarly, Reda maintained that a mandatory TDM exception along the lines of 'the right to read is the right to mine' would be the preferred tool to address the issue. Unfortunately, opposing positions often rely on a level of technical illiteracy, such as the claim that TDM is a highly specialised activity that can only be facilitated through cooperation (often at additional cost) with publishers. Another such claim is that it is easy to obtain a licence for TDM as you simply need to obtain it from the relevant publisher. This is clearly not the case as TDM can be performed on any material, not just the selected academic works held on the servers of big publishers. When dealing with widely distributed works, determining who holds the rights to every publication (and every element of that publication) then becomes not just very difficult but often near impossible. It is also highly unlikely that such rights holders would even have a TDM protocol in place.

Reda then noted that the technical illiteracy problem also extends to the notion that TDM puts strain on the publisher's servers and could for instance, slow emergency access needed by doctors in the midst of some or other medical procedure (presumably reading journals in the operating theatre). Such an exclusion is currently in place, and allows publishers to limit access to ensure the integrity of their servers. Although abuse of this exclusion is expressly forbidden under the InfoSoc Directive and Member States would be allowed to intervene under this directive, it would be difficult to prove that publishers' actions are unwarranted and Member States have been reticent to take action. Notably, the InfoSoc Directive also allows for rules to be overridden by contract, which is not the case in the new proposals.

Interestingly, Japan adopted a TDM exception in 2012 that covers both commercial and non-commercial TDM, with licencing costs being one of the main motivations for this move. Another interesting case was the District Court of Amsterdam which ruled that prohibiting TDM on the rights holder's side on the grounds of copyright, is a disproportionate impingement on the fundamental freedom of research. Significantly, this ruling indicates that there is a chance that the issue of TDM may be resolved through case law without any need for legislative intervention. Conversely though, this highlights the danger of placing narrow exceptions in to law, as the courts may then view anything outside of these narrow exceptions as contravening the law.

If however a legislative intervention is in fact needed, then such an intervention needs to go much further than the transitory reproduction right mentioned earlier. Instead of a limited TDM exception, it would be more effective to look at the scope of the reproduction right. This would then affect not only TDM but also other issues such as freedom of panorama. The key issue driving these concerns is that at its core, the digital revolution means that our perception of the world around us will increasingly be mediated through digital technology and whenever that happens copies will inevitably be made. As this technology then advances, new exceptions or legal changes are continuously needed, with powerful lobby groups regularly voicing opposition to such moves.

Against this background, Reda asked if it still makes sense to root copyright law intrinsically within the act of copying when this act is inherent to all digital interactions irrespective of the indented outcome. This leads to the situation where the act of copying is in itself potentially an infringement, instead of just the act of providing a copy to a third party. Of course in the offline world this is a logical approach to take and assists in enforcement, given that making multiple unsanctioned copies of a copyright work would facilitate distribution and infringement on the copyright holder's rights. In the digital world though, this is simply not the case as illegal copies are created as distribution takes place (not beforehand). The benefits gained from the analogue reproduction rights therefore no longer apply. Not only is this advantage lost but the limitation on reproduction interferes with the development of digital technology. This yet again indicates the need for a more profound reform, with the aim of bringing digital copyright legislation back into the balance struck by analogue copyright requirements. Specifically, something along the lines of 'the right to read is the right to mine' will free researchers to conduct research on material they already have legal access to without worrying about legal issues brought about through the use of digital aids.

Turning then to questions from the attendees, Julia Reda was asked what the best means would be by which an interested party could claim their work to fall under the TDM exception. To this Reda responded by stating that demonstrating a public interest is generally speaking, the best course of action. This would deny the need to show compliance with the other requirements and could easily be demonstrated by organisations which by definition have a public interest mission. Though she also noted that on this count the exception is in need of broadening to also include schools, as they should fit the bill even though they are not explicitly research organisations.

Addressing a point of uncertainty raised by an attendee, Reda further stated that a positive aspect regarding the Council's conclusion is that it refers to all bodies and organisations including citizen scientist and businesses, while none of these groups are clearly covered in the Commission proposal. Overall then, the Council can be seen as somewhat more progressive.

Lastly, a question was raised regarding the status of mining data which is being streamed, as such data would be transient in nature. Here Reda commented that if you were mining the stream in real time and not saving any of the stream's own data to your system, then you would be covered by the existing exemption. Though this is possible, it is not ideal. In this instance one would be forced to design IT systems on a compliance basis instead of focussing on what would be optimal from a TDM or technological perspective.

First panel: Copyright and database rights, data mining and data sharing.

The panel was chaired by Eleonora Rosati (Southampton) and included Estelle Derclaye (Nottingham, via Skype), Andres Guadamuz (Sussex), Trevor Callaghan (Google DeepMind), Carlo Scollo Lavizzari (International Association of Scientific, Technical, and Medical Publishers - Legal Counsel), and Margaret Haig (UK IPO).

After introducing the panel, Eleonora Rosati framed the key session topics with reference to the UK's introduction in 2014 of a TDM exception. The UK did so believing that this possibility was already allowed in the InfoSoc Directive. TDM is however still a heated point of discussion at the EU level, although the proposed EU exception differs from the UK option in both scope and approach. Starting then from a technological aspect crossing to the legal sphere, the first question was what the real potential of TDM is, and Rosati posed it to Trevor Callaghan.

Trevor Callaghan started by saying that the best way to think about TDM is to frame it against a problem. In this instance, it is a complexity problem brought on by our traditional ability to analyse data being far outstripped by the shear amount of data available. The increased use of technology to distil, segment and aggregate data in a meaningful manner, has reached a level of both speed and accuracy that is rapidly (if not already) becoming impossible for researchers to cope with without using tools such as TDM. The fundamental purpose of TDM therefore is to enable people to harness the power of the millions of pieces of information available to them.

As a follow-up question Rosati asked which sectors have the potential to benefit the most from TDM activities. Callaghan responded by referring to healthcare and medical research as being the preeminent areas to benefit from TDM activities. In this regard, new TDM technologies provide for lessons to be learned from data that is already available but which could not previous be inspected so rigorously. A second main area highlighted by Callaghan is computer science itself, where both the accretion of code and the complexity of code has led to a situation where complex systems are very difficult to debug. To this Carlo Scollo Lavizzari responded by stating that STM also views the future of TDM in this light and that the healthcare and pharmaceutical industries have been leading the way.

This discussion then led to a question from the audience on how one can guard against discriminatory biases creeping into TDM studies. Responding to this Julia Reda pointed to the example of predictive policing, where a certain community reports higher instances of criminality which leads police to commit more resources to this community. This action in turn increases the number of arrests and other actions associated with this community, which again leads predictive policing to commit even more resources to the eventual exclusion of other communities through a self-affirming bias. As a possible corrective measure, Reda then stated that TDM should move out of the niche academic position it currently occupies and become a cultural technique, but for that to happen people must be allowed to do TDM in the first place.

This led to a question regarding the IP issues connected to TDM activities, which was addressed by Estelle Derclaye. Starting with the *sui generis* right connected to TDM, Derclaye specifically discussed databases but also mentioned confidential information and trade secrets. The relevance of the *sui generis* right in connection to these databases lies not only in the protectable nature of data stored, but also in the fact that the definition of what constitutes a database is rather broad, meaning that protection is relatively easy to obtain. The *sui generis* right is subject to a number of exceptions, but these are however limited and optional, meaning that the Member States are under no obligation to implement them. One of these optional exceptions includes the use of data for education, research and non-commercial purposes. In addition to this exception being optional it also only applies to the act of TDM and not to the communication (publication) of results, for which a licence must be obtained from the rights holder. This is however, not the end of it as even if these exceptions have been implemented by a Member State and you could comply with all provisions, TDM may still be barred as any provision may be overridden by contractual agreement. In conclusion, Derclaye stated that this approach makes it near impossible to perform TDM in

the EU with regard to the *sui generis* right. A much better approach would be that previously proposed by Reda, such as either a fair use consideration or revisiting the reproduction right.

In response Carlo Scollo Lavizzari interjected that he could not wait to disagree with most of what Ms Reda had said, but there was one point of agreement (albeit from opposite ends): the proposed TDM exception is unnecessary and unwarranted. Scollo-Lavizzari then asked: if – for argument's sake - a broad TDM exclusion was imposed would one, for instance, be able to load the entirety of Google Maps and compare it to another source? Derclaye pointed out that Google Maps is not entirely a Google-owned product as Google has licences with other map producers and rights holders in different countries. Scollo Lavizzari then asked what the case would be if no such concerns existed. Trevor Callaghan responded, first, by stating that the core issue is that Google is both an intermediary and right holder. If one looks beyond this complication though, Google is positive about TDM as it is seen as a value driver. This principle, or starting point, is however often undone in practice due to the aforementioned complexities. On this point Andres Guadamuz commented how much one can already accomplish using Google's map APIs.

Scollo Lavizzari then commented that because STM views TDM as the main future vehicle for interaction with its publications, it will be publishing in an adapted manner to facilitate TDM activities – publishing content with TDM reading tools in mind to begin with. Interesting about this is that in Japan the TDM exclusion there is actually not applicable in the case of such adapted publications (original and non-original databases). The question then becomes what protection or other measures will be in place once TDM adapted publications become the norm. Considering this point, Derclaye commented that the *sui generis* right would remain unaffected given how broad the definition of 'database' currently is. Given the added value this may even make copyright protection more likely. To this Reda added that if the usability and accessibility of a database is so enhanced for TDM, people would be willing to pay for such access but that this would be an issue of market forces and not need copyright protections to enforce it. Specifically, this means using the rights holders' proprietary software over any other data mining tools.

At this point Ian Bourne (ICO) who would have been on the second panel, delivered a single presentation due to time constraints. Bourne's presentation focused on data protection, not just as data security but also as an enabling paradigm focusing on individual rights. Key to this is not only the ever-increasing amount of data collected on all citizens but also the increasing complexity of analytics and the potential for intrusion that it brings. In short then, if we have

more knowledge do we have less privacy and how do we manage the interaction between those two elements? In this regard the UK Anonymisation Network (UKAN) has been doing extensive research in to the science of privacy protection. There is a very specific need for this as privacy protection techniques have not entirely kept pace with advances in TDM and analytics.

A key and all too prominent example that Bourne presented is the issue of bulk data collection, which not only includes the activities of policing and security bodies but also advertising and other corporate concerns. The ICO as regulator, is tasked with ensuring that developments in the field of privacy protection keep pace with analytics and data collection advances. The issues concerned though, as is generally the case in the digital world, are no longer geographically limited, with data protection issues being addressed the world over. This gives rise to varying approaches and also varying understandings of what constitutes personally identifiable information. Thereby impacting on what can be protected under privacy regulations. One possible way of dealing with this issue is the ethical committee route pioneered in the medical field.

Lastly, Bourne addressed the potential of a possible public acceptability test. This could, and often does, yield different results to what the data protection community might expect. Thus data protection can be seen as having a mixed nature in both responsibilities and rights, underpinned by transparency. The latter can be difficult though when dealing with a public who varies from disinterested on one topic, to passionate on the next, while this level of engagement does not necessarily coincide with an increase in technical understanding. This notwithstanding, civic society interactions with these issues are increasing at a pace.

Returning to the UK's TDM exception and specific issues around its success rate, Eleonora Rosati introduced Margaret Haig from the UK IPO. As a point of departure, Haig referenced the 2011 Hargreaves Report and subsequent changes to the economic use of copyright. Broadly speaking, the recommendations in the report were based on exceptions taken from the InfoSoc Directive and other measures including the licensing of orphan works to unlock cultural heritage. It also includes improvements on the disability exception and various other aspects such as cultural preservation. Through this review process TDM was one of the issues that came up, though at the time it was primarily the research community asking for it. This was specifically in response to publishers asking for licence fees. Upon further investigation, it was found that a TDM exception would lead to efficiency savings of approximately £125m per year. Haig further highlighted the major benefits of the UK approach, including that

because it is for non-commercial purposes, instead of being limited to the user, any grouping from individuals to small teams and even public private partnerships, are allowed. This then also allows for commercial entities to do *pro bono* work.

After nearly two years the UK dispensation has yet to see any legal challenge brought and can be viewed as successful on this count. However, the IPO currently lacks sufficient insight into how researchers are using the TDM exception and a more nuanced assessment can therefore not be provided. What information is available though, mostly relates to medical research.

At this point Estelle Derclaye inquired as to why the UK did not introduce the same exception for database rights and if this was risk aversion based on the *sui generis* right. Haig confirmed that this could be partially the case.

Carlo Scollo Lavizzari then raised the issue of publishers charging for TDM and stated that at the time the main concern for publishers was not whether or not TDM would take place, but rather how it was done. He further stated that the main purpose of proposed contractual conditions was not monetary gain but rather control over infrastructure. By way of analogy, a publisher's database can be seen as a hotel with clients constantly coming and going, some to read and others to mine. The publishers would then present these clients with separate reading and mining rooms, allowing for data to be collected on usage. This is further facilitated through the APIs offered by the industry. The TDM exceptions that have been brought in though, have prevented publishers from exercising this level of control and has subsequently also barred them from gaining the associated insights. This highlights the need for a more dynamic regulatory environment as opposed to the imposition of rapidly outdated static rules.

Responding to these comments Julia Reda stated that the best way to encourage people to use the reading and TDM areas as requested would be to furnish them appropriately. This entails the provision of the appropriate TDM tools, leading to people using them voluntarily as opposed being forced to do so. Key here is the tendency of publishers to dictate to researchers, whilst the researchers themselves are surely in the best position to judge their own needs.

This discussion was then followed up by a question from the audience on the increasing use in TDM contractual obligations of what is referred to as a snippet. A snippet allows for the reuse of content limited to a particular length, often as little as 150 characters regardless of

the length of the work. In many fields, such as literature review for instance, such a requirement would be unworkable. This then leads to the question as to how these requirements relate to existing exceptions and how researchers are to deal with this. In answer to this query, Scollo Lavizzari stated that the concept of a snippet was introduced as a safety net to ensure that researchers were aware that some reuse was permitted. Furthermore, this requirement does not substitute the quotation exception and the latter therefore is the true determinant of allowed reuse length.

On this point, Andres Guadamuz commented that such contractual stipulations continuously obstruct or delay the process of research, as researchers often need legal reassurances. This is due to the TDM protection only being part of the issue, as the act of publishing research brings new concerns. In relation to the Commission's proposal, there are some notable questions though, such as the EU focus on the organisation as opposed to the UK's focus on the individual, as well as the mention of the non-commercial purpose in the UK version. It is also highly likely that the final version of the text may be significantly different from the current version. On this point Reda added that the Commission's best course of action would have been to write a mandatory research and education exception at the EU level which includes TDM and illustration for teaching, as this would have achieved the goal of simplification. Scollo Lavizzari, however objected to this notion on the basis that it would create a business model for large tech firms such as SAS and IBM, where they would pursue advances and deliver technologies in this field whilst excluding the rights holders. This while the rights holders have invested in building the databases being used.

Picking up on Scollo Lavizzari's points and referencing the report's accompanying Recital 10, Estelle Derclaye noted that research organisations should also benefit from the exception when they engage in public private partnerships. Furthermore, in Recital 13 the Commission acknowledges that performing TDM holds minimum harm to rights holders. Concluding therefore that there is no need for compensation to rights holders in relation to activities under the TDM exception.

Second Keynote: Madeleine Greenhalgh (Data Science Team, Government Digital Service, Cabinet Office.)

Madeleine Greenhalgh's presentation started with an overview of the Government Data Service's work on data science in government over the past three years. As a point of departure, Greenhalgh noted the Government Data Service's (GDS) realisation that current advances in technology and the accompanying increase in the amount of data available, both underscored the need for public service improvements and also provided significant opportunities for data mining. By way of example, the food standards agency has begun to use TDM on social media resources to predict Norovirus outbreaks ahead of lab reports. This provides real time data that is coming in one or two days earlier than would otherwise have been the case, which is viewed as invaluable in disease prevention and management. Along a similar vein, GDS has started to use service comments as predictors of demand. This entails the use of topic modelling to cluster prominent words and phrases indicative of high demand or service failures; thereby allowing for remedial action to be taken even before complaints are received via traditional channels.

Before the GDS or other bodies can embark on such projects though, heed must be paid to relevant legal requirements with relevant legislation including The Data Protection Act, Contract Law, The Computer Misuse Act, Intellectual Property Law, Copyright, Database Rights, Human Rights Act and The Digital Economy Bill. Added to this is the relevant guides and codes of practice which on the one hand aid in understanding, but on the other add to the sheer volume of requirements to be considered. This volume of relevant requirements and legislation can therefore present a significant obstacle to any new project. Once dealt with though, significant TDM opportunities are opened up within the current legislative framework.

One key consideration highlighted by Greenhalgh, was the issue of ethics. This of course not only includes those ethical practices enshrined in law, but also those not codified but still commonly held by society at large. A practical example of the latter would be the decrease in the use of corporal punishment even though it is still legal in the UK. Adding further complexity though is that notions of ethics are not only changing constantly, but are doing so at a pace faster than the law can hope to reflect.

The challenge then for the GDS is to bring together the relevant laws and best practice, enabling the civil service to more easily navigate the related issues. To facilitate this, a Data Science Ethical Framework has been built, with work continuing over the past two years up to publication in May 2016. This framework specifically references data mining and data sharing

by providing data scientists and their teams with robust principles. Acting not as fixed arbiters of what is ethically acceptable and what is not, but rather as a starting point for conversations on ethics in research. The six principles are as follows:

1. Start with clear user need and public benefit.
2. Use data and tools which have the minimal intrusion necessary.
3. Create robust data science models.
4. Be alert to public perceptions.
5. Be as open and accountable as possible.
6. Keep data secure.

The value of these principles comes from them being applied in a balanced manner as none outweighs the others. This notwithstanding, it is possible that a greater result in one area could mitigate a decreased result in another. For example, dealing with large scale viral outbreaks requiring a more intrusive data set.

Greenhalgh then went on to examine the case of web-scraping, which can be seen as the intersection of data sharing and data mining. By mining publicly available information from the web one is in fact simultaneously facilitating data sharing. An example of this in terms of governmental work is the Office of National Statistics (ONS) using web-scraping to gather data that feed into the CPI calculations. This has been made possible by an increasing number of retailers offering online shopping, thereby making price data available. A further implication is that data collection becomes easier, cheaper and more regular as the traditional physical and time constraints are removed.

These gains notwithstanding, there are notable ethical concerns to guard against. This is not only due to the more obvious issues of privacy but also because the act of data-scraping by definition facilitates data sharing. Because of this lack of a formal agreement on the extent and nature of the data sharing involved, the opportunity arises for the data to move beyond the sphere and use the original holder intended, even if this data originated in the public domain. This may of course be limited through the use of the robot exclusion protocol to bar scrapers from certain data. Unfortunately, this does not fully address the situation, especially in cases were the original holder of the data does not have control over the site were the data are stored or displayed. One recent example of this was the scraping of, and subsequent analysis on, user data from an online dating site. The results of the research were published along with the full profile information of the accounts scraped, which provoked a significant public backlash. In this case then, the data was in the public domain, but those who placed it

there did so for a specific purpose and did not expect an academic analysis to be performed on their personal details.

In conclusion, Greenhalgh presented the GDS's Data Science Ethical Framework's principles with regard to web scraping guidance:

1. Always respect the website terms and conditions & robots protocol.
2. Notify website owners of any plans to scrape their website on a large scale.
3. Schedule web scraping activities so as to minimise the impact on target websites.
4. Do not scrape websites anonymously – make sure an identifiable IP address is visible.
5. Obtain explicit agreement from the website owner for scraping a website for statistical purposes.
6. Ensure that any republishing data sourced from the web could not be interpreted as a breach of intellectual property rights.

Starting the question and answer session, an attendee noted that although an open and transparent approach is advocated here, there are no guarantees that researchers themselves will not hide the results of their research in distinct silos. In answer, Greenhalgh stated that this is indeed the case and that the subsequent accessibility of research findings constitutes an area which still needs improvement. The latter could of course also constitute a form of return on participation for website owners. As far as the public sector is concerned, the making available of such results or even the data used (if anonymised) should be conducted in terms of the framework previously presented.

When questioned about the challenges facing the public sector in this field, Greenhalgh commented that the framework is published but not mandated. As such, adoption largely rests on making departments aware of the framework and its applications. That notwithstanding, the Data Science Team have been active on this issue for more than three years and in that time have built up an understanding of the data science capacity and knowhow in other departments. Again, this understanding rests on reciprocity rather than regulatory requirements.

Lastly, with regards to a statement by Greenhalgh that 'the law affords leeway to the researcher, hence the need for ethical checks'; Sophie Stalla-Bourdillon asked if this was in reference to data protection or IP law. Greenhalgh stated that this was predominantly with reference to data protection, though it also refers to the interaction between various applicable pieces of legislation.

Second Panel: Data Protection, Data Ming and Data Sharing

The panel was chaired by Sophie Stalla-Bourdillon (Southampton) and included Tobias Koberg (Research Data Centre LlfBi), Christopher Brown (Jisc), Waltraut Kotschy (Data Protection Compliance Consulting, formerly Austrian DPA), Libby Bishop (Essex).

After introducing the panel, Sophie Stalla-Bourdillon presented the main aim of the panel as investigating the interaction between, and unique issues inherent in, privacy, data protection and data challenges. The first panellist to take on this discussion was Waltraut Kotschy, former head of the Austrian Data Protection Agency, who examined the relevant legal framework. Kotschy firstly noted that in Austria, legal discussion is constrained by reference to legal texts and as such, considerations not formally codified are not brought into the discussion. Working from this background then, she focused on the text of the new General Data Protection Regulation (EU-GDPR). In this regard, the implications of the new text will be varied dependant on different approaches in each national implementation. Moving away from any one national implementation though, it is possible to have a discussion of the text on its own merits.

The first point of note is that although TDM and data sharing are governed by a number of regulations at the EU level, none of these regulations actually define what exactly constitutes TDM and data sharing. Irrespective of the absence of such legal definition then, it can be taken that the term TDM suggest the analysis of a large amount of data in order to answer one or more specific questions. From a data protection point of view, it is interesting to examine whether or not these answers constitute data about data, or still possess the attribute of being personal data.

Data sharing, dealing with the making available of data to an outside party for another purpose, can be similarly interesting when viewed from a legal framework which is governed by the principle of purpose limitation. There is of course a longstanding exemption from the purpose limitation based on the principle of compatible or not incompatible purposes. Unfortunately, the Regulation does not offer exact guidance on the meaning of the term ‘compatible’ or ‘not incompatible’; however, in Art. 6(4) it names at least some criteria which are considered to be important for assessing whether a new purpose is “compatible”. It is therefore finally up to the practitioner to apply their own good judgement in the matter. There are however, certain purposes which are assumed to automatically meet the requirement of compatibility. These are historical and scientific research, statistical purposes and archiving purposes. What should further be born in mind here, is that these are exceptions from the standard legal position and

should therefore only be applied in a limited manner. Hence, these exceptions are not a leeway to take any course of action one wishes. Instead there should always be a focus on striking a balance, because these exceptions are generally aimed at serving important areas of public interest. These exceptions exist then to override protections where such actions are needed to serve a public good.

A further major requirement to bear in mind is the storage limitation principle. This is because you cannot have archiving as a legitimate form of processing data if you have a very strict storage limitation principle and no exception. This principle further presupposes that one is able to keep data for a long period of time; in a manner such that third parties are prevented from accessing said data other than for purposes permitted by the exceptions.

Turning to article 89(1) of the EU-GDPR, Kotschy noted that this article has an extremely long and varied history given different Council negotiations on the issue. In this respect, article 89(1) was initially exceedingly long, dealing with varying cases in depth, before disappearing altogether and then finally making a return in its current form. This latest incarnation though, is not very substantive. The article now only states the need for safeguards to be in place and then lists some possibilities, such as the principle of data minimisation. The problem with the latter though is that it is a general principle and would be present in all the stages of data collection and as such is not new here. One further specific instance mentioned in article 89(1) is the use of pseudonymisation, which holds that wherever pseudonymisation (or logically also: anonymisation) can be employed, it must be employed. There is however, one additional complication. Within article 9 there is a new provision which deals with the use of sensitive data for "*archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) based on Union or Member State law*". This formulation appears to suggest that either an additional Member State or EU law is needed to deal with the processing of sensitive data for archiving, research or statistical purposes. Considering the above, it is more than likely that the European Data Protection Board will have to provide additional guidance on the relevant issues.

In conclusion, Kotschy noted that from an Austrian perspective this new regulation is not that ground-breaking since the Austrian national implementation already has provisions dealing with these concerns. Furthermore, the Austrian provisions tend to be more precise in dealing with the consequences of specific actions.

Following from these points an audience question was fielded with regards to anonymisation frameworks. Regarding this, Kotschy commented that although the UK has a framework, The UK Anonymisation Network, there is no equivalent at the EU level which represents a significant need, especially when seen against the background of the new regulations. Joining the discussion, Libby Bishop added that there are other such frameworks that could be employed, but also the ONS's own framework which particularly deals with micro data. There is however, no generic framework that is suitable across all data formats and so longitudinal studies will need a different framework to cross-sectional studies, micro data being different from aggregate, etc. Stalla-Bourdillon then also added that part of the hindrance at the EU level rests on a lack of agreement on what constitutes anonymisation, with France opting for a definitive (irreversible) approach whereas the UK model takes a risk based approach.

Next, Tobias Koberg delivered a presentation on performing data protection in the practice of research. Starting at the most basic level, data protection has to consider first what type of data it is that you want to protect and from there which protection mechanisms are appropriate to employ. By way of example, Koberg referenced his own work at the National Educational Panel Study (NEPS) in Germany. NEPS collects longitudinal educational data but does not perform its own analytics, it instead provides the data to other national and international research bodies. The latter point is then particularly problematic from a data protection point of view.

The first points to note on data protection when dealing with the NEPS data include, that the study is voluntary, respondents provide written consent and are made aware of what data are being collected and for what purpose. The central problem then for the NEPS is not in the collection of the data, but rather in the facilitation of data sharing. Under the current German legislative framework, any data shared must first be anonymised though German Law also crucially includes a definition of what constitutes anonymised data and what level of anonymisation needs to be achieved. In this regard the German Law requires that there be a disproportionate investment in time and energy required to de-anonymise the data, when compared to any feasible gain from such de-anonymisation. In short then, for the requirement to be met de-anonymisation should not be worth the effort in the specific case being considered. In contrast the new regulation requires that one take into account all objective current elements, as well as future technological developments.

Staying with the present German dispensation, NEPS employs a portfolio approach to guide its actions in this field, with five different approaches constituting the portfolio. First there is the

organisational approach, where data are provided only to researchers with an association to a scientific institute. Second is the legal approach, which includes obtaining additional written assurances from the researchers with regards to data protection. Third is the informational approach, which includes significant user training and the provision of extra documentation. This is a key approach as it aids in ensuring that users operate within the parameters set by the NEPS, which is of specific importance as data protection breaches are normally accidental and not intended. Fourth, the technical approach is constituted of three different access modes which are the NEPS website, the study's remote access system (remote NEPS) and finally researchers can also work with the data on the NEPS premises. Fifth, there is the statically approach which changes data presentation and aggregation dependent on factors from the preceding approaches.

Returning to the differences between the German dispensation and the GDPR, Stalla-Bourdillon enquired if the new regulation is then seen as more restrictive. Koberg responded that there does not seem to be a clear path to compliance as the regulation requires one to account for future technological developments. This, on the face of it, appears to be an impossibility as one cannot account for technological developments that are five or more years in the future.

Next, Libby Bishop introduced the work of the UK Data Service. The service holds an extensive collection of social science (and linked fields) data and works to make that data available to researchers, whilst balancing protection with accessibility. This is facilitated via a 'five safes framework', which consists of safe people, projects, settings, outputs and data. Each of these 'safes' then has set actions to be taken, such as the training of researchers, the vetting of projects, where the data are held or used, reviewing outputs for disclosive properties and lastly to assess data safety. This latter element deals with anonymisation, which can be said to damage data quality as any distortion of the data will inherently affect its quality. It is therefore important to guard against heavy handed anonymisation whilst still recognising that it is an invaluable tool in data sharing.

When dealing with larger providers of data, such as the ONS, there is little to no need for additional anonymisation to be implemented, though this is sometimes not the case with smaller providers. The latter may experience issues around budgeting constraints and correct data handling techniques, which was one of the primary motivators for researcher training to be included in the 'five safes framework'. Outside of these instances though, the UK Data Service's main action around anonymisation is negotiation, in as much as during the intake of

data the completeness and accuracy of the anonymisation is discussed with the depositor and an agreement is reached. This discussion is then framed against the level access assigned to the data in terms of three tiers. These are open, which is public and has no registration; a midrange tier, which includes an end user licence and agreement to non-disclosure and no onward sharing as well as the provision of tracking information; lastly there is the secure tier where access to the data are either on-site in a secure room or off-site via a secure VPN with both these options allowing for the data to be analysed without the researcher ever getting a copy of the data.

Throughout all these considerations, anonymisation should be guided by the needs of the research project and the above framework, with safeguards such as researcher training increasing commensurate with an increase in the disclosiveness of the data. A concern in this regard though, is the ability to extract disclosive data from a seemingly anonymised data set by way of indirect identifier comparisons. On this point Stalla-Bourdillon enquired whether or not the UK Data Service viewed anonymised data as falling inside or outside the scope of UK data protection law. Bishop responded by stating that disclosive information would fall inside the regulations, but once anonymised it would fall outside

Christopher Brown (Jisc), the last panellist for the workshop, spoke on data sustainability and openness with Jisc subscribing to an ethos of open data and open access in its provision of digital solutions to UK based education and research. One of Brown's current projects is the research data discovery service, which aims to facilitate access to siloed data by collecting metadata and then storing said metadata in a central register. Issues encountered thus far include varying licencing requirements from participating institutions, institutions placing data behind logins so they can more easily extract their own metrics and also the harmonising of metadata collected from different schemas. This project has then also highlighted the need for data centres to harmonise their systems in terms of licencing, tagging and metadata. In addition to projects like these exposing inefficiencies in the operations of different data stores, it also allows for big data techniques to be brought to bear on the newly aggregated data so as to gain new insights.

From the point of view of the individual researcher, one problem that is often encountered is a lack of consideration for finer IP rights issues. This is due to the researchers being focused on gaining their own access to the data but not considering whether or not others will be able to replicate, access results or utilise any newly developed systems. A further possible

complication comes in the form of international projects where data from one country might be processed in another, which then significantly increases the regulatory burden.

Returning to the issue of sustainability, Brown noted that it is not merely about ensuring that projects are sufficiently funded, but is also concerned with the people and the policies associated with each project. One example of work done by Jisc in this regard was the use of the Janet network in the provision of secure data to bioinformatics researchers, where this application provided a controlled, safe and closed environment for data sharing and analytics to be performed.

Brown then also went on to highlight key issues in the training and support of data scientists and ensuring that all participants work to a recognised set of fair principles; namely that data should be findable, accessible, interoperable and re-useable. Additionally, it is proposed that adequate data stewardship should be made mandatory for all new research, with potentially 5% of the budget thus allocated, offering data stewards and experts training, as well as providing support tools for researchers to produce data management plans. Leading on from this discussion Stalla-Bourdillon then ask if we are to understand that IPR (Intellectual Property Rights) issues are more complicated and more difficult to address than data protections issues. Responding to this Brown stated that although he views both as solvable issues, it is the case that simple awareness of IPR issues amongst researchers is much lower than it is for data protection issues. Part of the issue relating to IPR is also the EU regulation, especially its focus on a research organisation which negates the possibility of citizen scientists performing TDM on materials normally available under the exceptions.

Sophie Stalla-Bourdillon then concluded proceedings by thanking all participants and guests and inviting them enjoy some post workshop refreshments.