Data Service Infrastructure for the Social Sciences and Humanities

EC FP7

Grant Agreement Number: 283646

Deliverable Report

Deliverable: D6.6
Deliverable Name: 6.6 Report about Preservation Policy-Rules (Preservation Challenges)
Deadline: 30 June 2014
Nature: R

Responsible: NSD, UiB
Work Package Leader: NSD

Contributing Partners and Editors: Trond Kvamme, NSD, Carla Parra, UiB, Koenraad De Smedt, UiB, Katrine Utaaker Segadal, NSD, Vigdis Kvalheim, NSD

Abstract

Use and reuse of research output based on personal information involves a potential conflict of interest between scientific progress and the improvement of the overall good for society and citizens and the risk of disregarding important data protection and privacy rights. In addition, issues related to data ownership and control (IPR and copyright issues) as well as the lack of appropriate and trusted research infrastructures for data and metadata preparation and sharing (digital deposit services), increasingly represent barriers to data preservation and data sharing. These challenges are further reinforced by a fragmented set of data protection and IPR laws, and a distributed data preservation and curation infrastructure.

In this report we will identify and discuss the problems that may occur when preserving SSH data in the emerging European data preservation e-infrastructure. We consider the proposed EU General Data Protection Regulation (GDPR) and the ongoing modernisation of the EU copyright framework as key aspects of the emerging e-infrastructure, as they will have severe impact on the development of new policies and policy-rules for researchers and data curators.

After a general introduction on the background, objectives and methodology of this report (Section 1) we consider the proposed EU General Data Protection Regulation (Section 2) by analysing key articles in the proposed GDPR. We also look at Intellectual Property Rights (IPR) and copyright issues in preservation and sharing by focusing on the extent of copyright as well as possible exceptions in selected countries, the ongoing European reform, and various licensing schemes (Section 3).

In Section 4 we draw some conclusions and provide a set of general considerations and recommendations for preservation policies that can work as a guide for setting up and maintaining a trustworthy data preservation and access environment.

Content

1 Introduction

1.1 Overall objectives of work package 6

This report is part of Work Package 6 "Legal and Ethical Issues" of the Data Service Infrastructures for the Social Sciences and Humanities" (DASISH) project. The focus of WP6 is on legal and ethical issues for the collection, curation, preservation and dissemination of data in the social science and humanities (SSH) area. The objectives of the work package, according to the Description of Work (DoW) of the DASISH project, are to identify legal and ethical issues, constraints and requirements for all data types occurring in the SSH domain; to cope with the legal and ethical challenges imposed by new data types emerging in the social sciences; and to look for professional long-run infrastructure preservation strategies for data in the SSH.

1.2 Objectives, background and focus of task 6.3

In this task we are looking at ethical and legal issues and challenges that confront researchers, data owners and digital repositories in the emerging European data preservation infrastructure environment. Building on Deliverable 6.1 ("Report about new IPR challenges") and Work Package 4 ("Data Archiving"), the main focus is on data preservation and data sharing, notably issues related to data protection, data ownership and copyright.

Managing privacy and access is an issue of major importance and concern in the current infrastructure landscape, and various procedures to ensure the optimal balance between data protection and data access are being developed and tested among various data producers and research fields across Europe. A vast majority of these models are based on technical solutions (disclosure techniques, remote access and execution models) to protect privacy in the accession process. In this report we identify the problems that may occur in relation to preservation and sharing of sensitive data, and the policy-rules that need to be considered when sensitive data will be preserved in a distributed data preservation and curation infrastructure. Further, we will identify and define policies and policy- rule mechanisms that guide the preservation and access rights while maintaining trust. Data protection and copyright issues has a major influence on the possibilities for long-term preservation and data sharing, and we consider the proposed EU General Data Protection Regulation (GDPR) and the ongoing modernisation of the EU copyright framework as key aspects of the emerging infrastructure landscape. Both will have considerable impact on the development of new policies and policy-rules in digital repositories.

The increased focus on legal and ethical issues, constraints and requirements for all data types in the SSH domain occur as a result of the development of new and powerful tools and methods for data mining, and the integration and linkage of multiple data sources. The challenges imposed by these changes are connected to the growing importance of data in most people's lives. The collection, storage, and analysis of data is on an upward growing trajectory, where the declining cost of collection, storage, and processing of data seems to grow exponentially with the amount of data. This combined with more recent sources of data like sensors, cameras, geospatial and other observational technologies ("internet of things"), means that we now "…live in a world of near-ubiquitous data collection"[1]. The term "big data" reflects this growing technological ability to capture, aggregate, and process an ever-greater volume, velocity, and variety of data ("the three

[1] Executive Office of the President (2014): *Big data: seizing opportunities, preserving values:* http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

V's")[2]. According to some sources, the World Wide Web now (as of May 2014) contains somewhere between 2.2 and 2.5 billion webpages[3] and almost 700 exabytes of accessible data[4] (that is 700,000,000,000 gigabytes).

Research using 'big data' is of growing importance for several research disciplines in Europe. Increasingly, scholars of arts and humanities, information science, linguistics and social studies do their research based on language material from all kinds of audio-visual content (films, TV series, music, speech, etc.), e-books, magazines, journals, newspapers and different types of user-generated content. New forms of data on human activities are now being recorded in a variety of domains from blogs and social media to new forms of data which arise from digital processes involving registrations, transactions, sensing devices, internet activity, telecommunications, retail sales, utility consumption, etc. These data types are not necessarily designed for research, but they may have significant research value, especially when linked across domains, or to survey data.

The fusion of different kinds of data, e.g. the linking of survey data with administrative data involves a risk of disclosure of personal data. Similarly, integrating diverse data can lead to what some analysts call the "mosaic effect," where the new dataset is much richer in detail than the individual dataset. From this new dataset personally identifiable information can be derived or inferred (although the new dataset do not include any direct personal identifiers), bringing into focus a picture of who an individual is and what he or she likes[5].

This study focuses on the ethical and legal issues that may occur in the *long-term preservation* of these new digital resources. We interpret long-term preservation in a broad sense, namely as the process(es) that "…refer to policies, strategies and actions that ensure permanent access to digital content over time"[6]. This broader understanding of preservation can include the transfer of data from point A (e.g. a researcher) to an archive or a data deposit service (ingest); the data storage and access arrangements (long-term preservation, or digest); and the dissemination (sharing, reuse) phase. Maintaining access rights (i.e. reuse of data) is a key function of research data repositories and a challenge for long-term data preservation. In this context we highlight these challenges by analysing key articles in the proposed GDPR. We also look at Intellectual Property Rights (IPR) and copyright issues in preservation and sharing by focusing on the ongoing European reform and various data licensing schemes.

The articulation of the legal text in the GDPR and the ongoing IPR reform will have significant impact on the definition of policy-rules for SSH data infrastructures now and in the future, and by exploring these issues we aim to look for professional long-run preservation strategies based on e-Infrastructures for data in the social sciences and humanities.

1.2.1 Utility and risk in research and data preservation

In the eighteenth century the philosopher Immanuel Kant introduced the Humanity Formula, which states that we (humans) should never act in such a way that we treat Humanity, whether in ourselves or in others, as a means only but always as an end in itself [7]. That is, one should always treat rational beings as having intrinsic value or worth, not as mere instruments or objects having only extrinsic

[2] Ibid.
[3] The size of the World Wide Web: http://www.worldwidewebsize.com/
[4] Facts Hunt: http://www.factshunt.com/2014/01/total-number-of-websites-size-of.html. According to the same site the full size of data (including *inaccessible* data) is over 1 yottabyte (10^{24} bytes).
[5] Executive Office of the President (2014): *Big Data and Privacy: A Technological Perspective:* http://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf
[6] EC: Survey on scientific information in the digital age: http://ec.europa.eu/research/science-society/document_library/pdf_06/survey-on-scientific-information-digital-age_en.pdf
[7] Stanford Encyclopedia of Philosophy: Kant's Moral Philosophy: http://plato.stanford.edu/entries/kant-moral/

value. Kantianism also holds that moral standards should be universal: moral principles are rules that would be followed by all rational beings with a good will.

Later, in the nineteenth century, Jeremy Bentham and John Stuart Mill articulated the concept of *utilitarianism*. In the utilitarian world view the right action is understood entirely in terms of consequences produced, and the 'best' outcome is that which maximizes the overall good (or utility) for all people in the long run. That is, it seeks to bring 'the greatest amount of good for the greatest number of people'[8].

Research that involves humans often involves a potential conflict of interest between scientific progress and the improvement of the overall good for society and citizens (utility), and the risk of treating research subjects as mere means to an end by disregarding potential risks of doing harm to the individual involved in the research.

> «Privacy is the most comprehensive of all rights and the right most cherished by citizens of a free nation.» (Justice Louis Brandeis, 1928)
>
> «A people who mean to be their own Governours, must arm themselves with the power to which knowledge gives.» (James Madison, 1822)

These two quotations highlight an important dilemma in the modern discussion on privacy and open access to information – the obligation to protect the sanctity of privacy, while at the same time ensuring a basis of information for society and democratic access to data. Moreover, there is recognition that a shift in the balance between the appreciations of different values can generate consequences which impinge on vital democratic values.

The idea of privacy or the individuals` right to protect his or her integrity and to be let alone has a long history and the system of norms which are the foundations for modern privacy regulations and ethical guidelines emerges from these traditions. The advantage of modern data protection regulations and other statutory regulations in this area is that privacy is being protected by law to a greater extent than earlier. Our social sense of what is right and wrong has been formalised and institutionalised; the social sense has, via ethical norms, become legislation. Further, modern privacy regulations apply to information security whereas the traditional focus was on protection from unwanted intrusion and physical and psychological harm. Consequently, anonymisation and various disclosure techniques are prescribed as the most important tool for protection in the modern information and communication society.

This potential conflict of interest between the benefit of knowledge – knowledge as a goal in itself – and the moral duty to reflect upon the proper way to gain that knowledge can take place between several different actors involved in the data lifecycle. Here we will focus on two areas of potential friction.

The first area involves the possible friction of interest between data creators, disseminators and curators on the one hand, and the research subjects (society) on the other. Here, challenges to privacy may arise because large amounts of data are collected and analysed so that personally identifiable information can be derived or inferred from the results, bringing into focus a picture of who an individual is and what he or she likes. These challenges may be compounded by limitations of traditional techniques or technologies used to protect privacy (disclosure control techniques) or lack

[8] Stanford Encyclopedia of Philosophy: The History of Utilitarianism: http://plato.stanford.edu/entries/utilitarianism-history/

of attentiveness or knowledge of data protection and privacy issues among researchers, disseminators and curators. On the utility side, there is scientific progress, breakthroughs and innovative solutions with possible beneficial outcomes for data users, citizens and society in general, now and in the future.

A second area concerns the potential conflict of interest between data creators (researchers), disseminators and curators themselves, especially in the process of sharing and reusing data. Possible issues of risks and concern include ownership to data (IPR and copyright issues), concerns about researchers' freeriding on data gathered by other researchers, fear of losing control over data, lack of proper data sharing infrastructures or lack of standards for sharing and preparing metadata and data. On the utility side there are all the benefits connected to sharing, preserving and reusing data. Sharing data may encourage scientific enquiry and debate, and encourage the improvement and validation of research methods. It may enable scrutiny of research outcomes and facilitate research beyond the scope of the original research; it may increase the impact and visibility of research and lead to new collaborations between data users and data creators; and it may provide important resources for education and training, for other researchers and for society and citizens in general[9]. Preserving data with professional disseminators and curators (data archives, repositories or data centres) can for example ensure the safe-keeping of research data in a secure environment and provide long-term preservation and back-up of data, while enhancing the visibility of data and enabling more use and citation.

All these issues and examples of possible conflicts between *utility* and *risk* involve considerations concerning *ethics* and *law* and are closely connected to and intertwined with the rapidly evolving technological platforms and research methods that are being applied in generating and sharing data.

1.2.2 Legal issues and ethical conduct in science

As a concept, 'research ethics' refers to a complex set of values, standards and institutional schemes that help constitute and regulate scientific activity. Ultimately, research ethics is a codification of ethics of science in practice. In other words, it is based on general ethics of science, just as general ethics is based on common-sense morality[10].

When research involves obtaining data from people, researchers are expected to maintain high ethical standards such as those recommended by professional bodies, institutions and funding organizations, both during research and when sharing data. With regard to ethical guidelines and codes of ethics there is no shortage of resources[11].

The guidelines are tools for researchers and data curators. They identify relevant factors that should or ought to be taken into account, but which must often be weighed against each other, as well as against other important considerations[12]. Some of the most common ethical standards embodied in various guidelines can also be found in legislation. As such, in many ways legislation and research ethics overlap. For example, in many countries the collection, processing and preservation of

[9] UK Data Archive, on the benefits of depositing and sharing research data: http://www.data-archive.ac.uk/deposit/why

[10] National Committees for Research Ethics in Norway: Guidelines for Research Ethics in the Social Sciences, Law and the Humanities: https://www.etikkom.no/en/In-English/Committee-for-Research-Ethics-in-the-Social-Sciences-and-the-Humanities/

[11] For a general overview and discussion of various ethical guidelines and frameworks see DASISH D6.1: Legal and ethical issues: http://dasish.eu/publications/projectreports/D6.1_final.pdf

[12] National Committees for Research Ethics in Norway (See note above)

research data may have a legal requirement for consent on the part of those who actively participate in a research project. This is also an important ethical consideration.

Like ethics in general, research ethics embraces both personal (e.g. the researcher) and institutional (e.g. the archive or repository) morality. As such, the obligation to respect research ethics applies to research in general. Individual researchers, project managers, research institutions and the appropriating authorities all share the responsibility.

The types of data that have the potential to harm and infringe on the research subjects' personal rights, are often a vital resource for academic research across a wide range of disciplines. Such data underpin observational, often longitudinal, studies and have led to significant advances that might have been otherwise impossible[13].

In the ***Social Sciences***, for example, studies using personal data have produced invaluable insights into a wide variety of socio-economic factors, opinions and behavioural patterns of a wide variety of individuals. Results have often led to an evidence base at the disposal of policy makers to address key societal challenges in Europe today. Producing key evidence of this type would become much more difficult without the appropriate provisions for scientific research in data protection regulations. In fact, some research depends on access to personal data and the statistics derived from personal data, like for instance the study of whether government policies have been effective and how they could be improved. Increasingly, researchers are seeking to link together administrative information about one individual across a range of sectors – such as health, education and welfare – to build a better picture of how these complex interactions affect citizens' lives and wellbeing.

In the ***Humanities***, scientific research also relies heavily on the collection, retrieval and analysis of personal data. Examples include: work on language diversity based on speech recordings; studies of cultural innovation, using an interaction design involving individuals and groups, and research on historical transformations based on archival material such as letters, diaries, family photographs and other personal visuals. Research of this type may have contributed to a better understanding of the social transformations and of the processes of cultural and social identity formation which underlie these.

Examples in ***Medical Studies*** (including studies of Life Sciences) include work that has demonstrated the long-term value of drug interventions, compared hospital death rates, and uncovered the role of genetics and environment in disease. Results have led to more effective treatments for many diseases, including chronic diseases. Moreover, these studies have also contributed to providing scientific evidence for policies that help make health-care systems more efficient and less costly. As such they have enhanced the quality of life for citizens. Discovery of new phenomena in medicine may occur only after going back to the original patient files and stratifying them according to the new variables. This type of analysis – which would not have been possible under data protection rules that prevent re-linking data and individuals – has for example led to the identification of new susceptibility genes for diabetes or new groups of patients that present different outcomes after breast cancer.

[13] This statement and the subsequent examples are mainly taken from: *Science Europe: Position statement on the proposed European GDPR*: http://www.scienceeurope.org/uploads/Public%20documents%20and%20speeches/SE_DPR_Position_FIN.pdf

1.2.3 Preservation, access and data sharing challenges

The examples from different SSH disciplines above are closely connected to challenges concerning the preservation, access to, and sharing of research data. The UK Data Archive (UKDA) identifies five key principles of research ethics that have a bearing on *sharing or archiving* sensitive research data[14]:

- A duty of *confidentiality* towards informants and participants.
- A duty to *protect participants* from harm, by not disclosing sensitive information.
- A duty to treat participants as intelligent beings, able to make their own decisions on how the information they provide can be used, shared and made public (through *informed consent*).
- A duty to *inform participants* how information and data obtained will be used, processed, shared, disposed of, prior to obtaining consent.
- A duty to wider society to *make available resources* produced by researchers with public funds (data sharing required by research funders).

However, several studies have shown that researchers still find barriers to sharing and archiving of their data. DAMVAD[15], Tenopir (2011)[16], the European Commission (2012)[17], and the Parse-Insight[18] project all confirm that many researchers are still undecided on the issue of sharing data. It seems that many researchers find sharing and archiving to be a difficult and complex issue and lack of incentives seems to be one of the central barriers to sharing. Time for preparation and lack of infrastructure are other barriers. The DAMVAD study finds that the barriers to sharing can be divided into three main categories:

Legal

- Privacy concerns.
- Shared ownership to data (IPR and copyright issues).
- Lack of knowledge on legal issues related to data.

Sociological

- Lack of incentives/credit to researcher.
- Concerns about researchers' freeriding on data gathered by other researchers.
- Fear of losing control over data.
- Fear of losing 'scientific edge'.
- Fear that others might not understand data.

Technical

- Lack of infrastructure.
- Sharing data is time-consuming.

[14] UKDA: Ethical/legal overview: http://www.data-archive.ac.uk/create-manage/consent-ethics/legal

[15] DAMVAD: *Sharing and archiving of publicly funded research data - Report to the Research Council of Norway*

[16] Tenopir, et.al. (2011): *Data Sharing by Scientists: Practices and Perceptions*
http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0021101

[17] European Commission (2012): Online survey on scientific information in the digital age
http://ec.europa.eu/research/science-society/document_library/pdf_06/survey-on-scientific-information-digital-age_en.pdf

[18] Parse Insight (2009): *Insight into digital preservation of research output in Europe, survey report*
http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf

- Lack of standards for sharing and preparing metadata.
- Lack of technical skills.

Hence, archiving and sharing of data involves a number of technical, financial, legal and ethical obstacles. While overall legal/ethical guidelines and stakeholder policy goals on data preservation are agreed upon, many questions still stand in the way of effective and successful implementation of the principles of trustworthy long-term preservation and accessibility of research data.

1.3 Methodology and further outline of the report

This report is mainly focusing on the *legal* issues that are emerging in the interaction between new technologies and new data types. We have chosen to emphasise the legal framework through an understanding of a legal regulation as a formalisation of ethical norms. A legal framework is an extension of an underlying ethical framework; where the ethics provides advice and guidelines, the law provides stronger degree of protection of specific norms, values and interests that society agrees upon through a formalised set of rules. In our case this involves values and interests that need special protection, such as the interests in privacy and protection of personal data, weighted against the needs of researchers to access to information and knowledge.

The problems and obstacles that may occur when preserving SSH data in the emerging European data preservation and infrastructure environment can only be solved by an attentive focus on the emerging European legal framework. The rest of the report consists of two main segments. In the first segment, we look at the recent developments of the General Data Protection Regulation (GDPR) of the European Union and the debate concerning the effects on preservation and sharing of research data. We focus on selected aspects of the regulation that may affect the processing and long-term preservation, access and reuse of research data. We discuss some of the changes that have occurred in the legal text from the original Commission proposal of 2012 to the final proposal that was voted for and supported by the Civil Liberties, Justice and Home Affairs Committee (the LIBE Committee) and the EU Parliament in 2014 and the possible implications these may have for data archives and repositories.

The second segment of the report is centred on Intellectual Property Rights (IPR) and the challenges posed by copyrighted data. By looking at copyright exceptions in a selection of countries, the ongoing European copyright reform, and a selection of licensing schemes, we will highlight some of the most important copyright issues that should be considered when creating and implementing preservation policies. In many cases, researchers in the social sciences and humanities share their data with other researchers, either by putting the data in the public domain, e.g. in a research data archive, or by retaining copyright by licensing the data, often without financial compensation, but sometimes with an embargo period. When data are archived in repositories, these repositories have to take copyright and licensing conditions into account.

For both segments we utilize a combination of literature review and 'case study' exploration to gather descriptive information about relevant issues. For the GDPR we highlight the amendments and articles in the regulation that are most relevant and that may influence the SSH research community; the discussion of the amendments is then thematically arranged into sections. The development of the GDPR from the 2012 Commission proposal, through the revised Albrecht report and the final LIBE proposal that was voted for in 2014, works as framework for the discussion of data protection issues. Central to the discussion are the official statements from various research

organizations. Recent legal developments in the EU also work as a backdrop for the discussion of IPR issues.

We conducted a literature review of relevant issues with a main focus on academic literature, reports from similar and related research and data infrastructure projects, specific disciplinary and research materials, expert statements and relevant websites. Based on these resources, we identified a set of legal and ethical issues relevant to long-term data preservation. This information was then coordinated and consolidated to formulate a set of general recommendations concerning policy-rules that guide the preservation and access rights while maintaining trust.

2 The General Data Protection Regulation (GDPR)

2.1 Introduction

Currently, personal and sensitive data in the European Union are protected by domestic implementations of the Data Protection Directive (95/46/EC)[19]. The basic principles, ensuring a functioning internal market and an effective protection of the fundamental right of individuals to data protection, are even more valid today than they were almost 20 years ago. The Directive establishes that personal data is any information relating to an individual, and that it applies when a person can be identified, directly or indirectly. The problem with the Directive are the differences in the way that each EU country implements the law, and how these have led to an uneven level of protection for personal data, depending on where an individual lives or where a researcher processes or preserves his data.

Consistent with the advisory nature of an EU directive, the member state data laws vary widely. While local laws offer data subjects at least the Directive's core protections, some add extra rights. Moreover, all member states have created their own unique Data Protection Authorities, compliance structures, notification processes, and other bureaucratic procedures. In short, questions about how to comply with data laws in Europe usually end up at the member state, as opposed to EU, level[20].

As the Directive has failed to achieve a proper harmonization, giving individuals, companies and researchers differences in data protection requirements, it has become increasingly difficult for different stakeholders to abide to the different data protection implementations. This lack of legal harmonization in data access regimes due to legal uncertainty leaves several unsolved issues for researchers. In fact, gaps, inconsistencies and contradictions may turn up when researchers are involved in cross-country research and data sharing.

For example, in the UK anonymisation is defined as "…the process of turning data into a form which does not identify individuals and where identification is *not likely to take place*"[21]. In Germany, there is a more specific description of anonymisation. The *Bundesdatenschutzgesetz* (the German Federal Data Protection Act) states that: "Rendering anonymous means the modification of personal data so that the information concerning personal or material circumstances can *no longer or only with a*

[19] DIRECTIVE 95/46/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:PDF

[20] White & Case: International Data Protection and Privacy Law

[21] UK Information Commissioner's Office: What is anonymisation?: http://ico.org.uk/for_organisations/data_protection/topic_guides/anonymisation

disproportionate amount of time, expense and labour be attributed to an identified or identifiable individual.[22]"

At the same time, since the adoption of the Directive a lot has changed in the area of data protection, "…notably technological developments, increased collection and processing of personal data, including for law enforcement purposes, with a patchwork of applicable data protection rules and globalization of markets and cooperation"[23].

The aim of the recent data protection reform in EU has been to modernize the principles from the 1995 Data Protection Directive and to strengthen citizens' rights and thereby help restore trust. Better data protection rules are aimed at the EU citizens so that they can be more confident about how their personal data is treated, particularly online. The new rules "…will put citizens back in control of their data"[24].

This recent development may affect the balance between the needs of the citizens and society (protection from *risk*) on the one hand, and the needs of the researchers, curators, disseminators and society (*utility* and progress) on the other hand. In the following subsection we will discuss some of the proposed legal amendments in the new General Data Protection Regulation (GDPR), highlighting how the proposed regulation represents a shift of balance between the protection of the two important societal values - information privacy and access to information - will affect the processing, preservation and sharing of research output in the SSH research communities.

2.2 Background and recent developments

On January 25, 2012 the European Commission published its proposal to reform the European Union's legal framework with regard to the protection and processing of personal data. The proposal includes two different legal instruments: a General Data Protection Regulation[25] covering data processing by the private sector and public authorities, and a General Data Protection Directive applicable to law enforcement[26]. The former is most relevant for our purposes and will be discussed in more detail in the following. The regulation[27] is set to replace the current Data Protection Directive 95/46/EC.

For research communities, the main question has been to what extent the proposed regulation would create safe and predictable conditions for research activities. The general view within the

[22] German Federal Data Protection Act, Section 3(6): http://www.gesetze-im-internet.de/englisch_bdsg/englisch_bdsg.html#p0028

[23] See explanatory statements in LIBE draft report 2012/0011 (COD) dated December 17, 2012. http://www.europarl.europa.eu/meetdocs/2009_2014/documents/libe/pr/922/922387/922387en.pdf

[24] European Commission - MEMO/14/186, 12/03/2014: http://europa.eu/rapid/press-release_MEMO-14-186_en.htm

[25] COM(2012) 11 final. 2012/0011 (COD): *Proposal for a Regulation of the European Parliament and the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)*. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2012:0011:FIN:EN:PDF http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2012:0011:FIN:EN:HTML

[26] COM(2012) 10 final. 2012/0010 (COD): *Proposal for a Directive of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data by competent authorities for the purposes of prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and the free movement of such data*. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2012:0010:FIN:EN:PDF http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2012:0010:FIN:EN:HTML

[27] "Regulations are the most direct form of EU law - as soon as they are passed, they have binding legal force throughout every Member State, on a par with national laws. National governments do not have to take action themselves to implement EU regulations. They are different from directives, which are addressed to national authorities, who must then take action to make them part of national law." http://ec.europa.eu/eu_law/introduction/what_regulation_en.htm

scientific communities in Europe has been that for the most part the regulation has contributed to more stable conditions. In certain areas, research interests are highlighted and strengthened. for example by clarifying that the different legal grounds for processing personal data are put on an equal footing and that it is legal to process non-sensitive personal data for research purposes without consent and without a balanced assessment of interests being required. In other areas, research interests are somewhat weakened, for example by dropping the special provision stating that subsequent processing for research purposes is not incompatible with the original purposes.

On December 17, 2012, the main rapporteur for the GDPR in the EU Parliament, Jan Philipp Albrecht, issued a draft report on the GDPR for the EU Parliament's Committee on Civil Liberties, Justice and Home Affairs (the LIBE Committee)[28]. The report expressed legitimate concerns regarding the growth of marketing interests that are challenging personal data protection and privacy. As a result, the report proposed substantive amendments to the Regulation and in the succeeding period many stakeholders in both the private and public sectors began to issue papers reacting to the proposal[29]. The report was a source of widespread concern, especially in the health science communities in Europe. Several academic organisations expressed strong concerns that the new version had gone 'too far' and put severe restrictions on the processing and preservation of research data. The reason for this was that the Albrecht Report proposed several amendments to the Commission's proposal. Among the most important changes were the deletion of several research exemptions and provisions that initially had been included in the Data Protection Directive 95/46/EC (and was continued in the proposal from the Commission) to highlight the research sector's legitimate need to process personal data, and to ensure a balance between research progress and data protection. For the research community it has therefore been seen as crucial that these exemption provisions are continued, so that the framework conditions for research can be maintained or, if possible, improved.

Based on input and draft opinions from stakeholders and other committees in the EU Parliament (including the Employment and Social Affairs Committee; the Industry, Research and Energy Committee; the Internal Market and Consumer Protection Committee; and the Legal Affairs Committee[30]), on November 21, 2013, the LIBE committee voted to approve a new compromise Draft Regulation[31]. On March 12, 2014, the European Parliament accepted the proposal, including the amendments proposed by the LIBE Committee[32]. The final Regulation adopted by the European Parliament retracted some of the amendments and adjustments suggested by the Albrecht Report.

There is still some uncertainty connected to the further processing of the Regulation, but the European Council aims for its adoption in late 2014 and the regulation is planned to take effect after a transition period of approximately two years. To become law the proposed Regulation has to be adopted by the Council of Ministers using the "ordinary legislative procedure" (co-decision). Once the Council has reached agreement on the text of the proposed Regulation, a 'trilogue' between the

[28] LIBE draft report 2012/0011 (COD) dated December 17, 2012. Officially presented during a LIBE meeting on January 10, 2013. http://www.europarl.europa.eu/meetdocs/2009_2014/documents/libe/pr/922/922387/922387en.pdf

[29] Selected stakeholder' papers can be found at the WSGR EU Data Protection Regulation Observatory: http://www.wsgr.com/eudataregulation/stakeholders-position-papers.htm#

[30] For a list of involved committees and their statements, see: http://www.wsgr.com/eudataregulation/process-updates.htm

[31] COM(2012)0011 – C7-0025/2012 – 2012/0011(COD): Ordinary legislative procedure: first reading. http://www.europarl.europa.eu/sides/getDoc.do?type=REPORT&mode=XML&reference=A7-2013-0402&language=EN

[32] P7_TA(2014)0212: Protection of individuals with regard to the processing of personal data: http://www.europarl.europa.eu/sides/getDoc.do?type=TA&reference=P7-TA-2014-0212&language=EN

Parliament, the Council and the European Commission will be established to agree on the final text. Following the trilogue, the proposed Regulation will be put to a vote of the Parliament and, if adopted, there will be a final implementation period before the Regulation comes into force in the EU Member States. On June 6, 2014, the European Commission (by the EU Commissioner for Justice) confirmed that the data protection reform is on track to ensure "the completion of the Digital Single Market by 2015"[33].

In the following subsections we will highlight the key articles in the Regulation that may influence on the balance between the protections of the individual on the one hand, and the processing, preservation and sharing of research data on the other hand. To highlight these issues we will in many instances compare the content of the original Commission Proposal of 2012 (from now on shortened to 'COM') with the final text that was adopted by the LIBE Committee in the European Parliament in 2014 (from now on shortened to 'LIBE'). A tabulated comparison of legal text can be found in Appendix 1.

2.3 Important provisions for the research sector

2.3.1 Definitions and scope

Personal data

The definition of 'personal data' is one of the key concepts for the protection of individuals by the current EU data protection instruments and triggers the application of the obligations incumbent upon data controllers and data processors[34]. The definition is laid out in Articles 4(2), Article 9(1) and Recital 23.

The original Commission proposal definition of personal data was separated into two sections and included a definition of the 'data subject', which is said to be a person who can be identified, directly or indirectly, "*…by means reasonably likely to be used by the controller or by any other natural or legal person"*. The inclusion of the 'reasonably likely' specification in the COM proposal was a liberalization of the definition and scope of the 'personal data' concept from the 95 Directive, in the sense that the 'reasonably likely' concept, at least theoretically, might limit the scope and volume of cases that are to be processed by Data Protection Authorities (DPA).

The test of 'means reasonably likely to be used' has been removed from the final LIBE version and is instead included in Recital 23. This might influence on the proportionality of determining whether data can lead to the identification of an individual. Hence, the LIBE Regulation, interpreted in its strictest sense, might mean that all countries must take into consideration even the *theoretical* possibility of personal identification. A theoretical possibility of identification lowers the threshold for considering information as directly or indirectly identifiable. The Information Commissioner's Office (ICO) in the UK expresses concerns that "…the mere possibility of identification widens the scope of personal data too much" and "…this could have a chilling effect in respect of the release of anonymised information derived from personal data, for example under freedom of information law"[35]. The ICO thinks that the test of actual identification or the reasonable likelihood of

[33] Press release from Justice Council: http://europa.eu/rapid/press-release_SPEECH-14-431_en.htm
[34] COM(2010) 609 final: Communication from the Commission to the European Parliament, the Council, the Economic and Social Committee and the Committee of the Regions: http://ec.europa.eu/justice/news/consulting_public/0006/com_2010_609_en.pdf
[35] UK Information Commissioner's Office: Comparative analysis of the European Commission text and the European Parliament's LIBE (civil liberties) Committee amendments http://ico.org.uk/news/blog/2013/~/media/documents/library/Data_Protection/Research_and_reports/Proposed-draft-EU-General-Data-Regulation-and-law-enforcement-Directive-20140124.pdf

identification (as expressed in the original COM proposal) should be maintained as it has worked well in practice in the UK and proved useful in assessing borderline cases. In their view it has been effective in ruling information outside the scope of data protection law where identification is not reasonably likely.

Although the statement is maintained in Recital 23, the removal of the segment from the *legal* text implicitly broadens the scope of data protection law to include all cases where the mere *possibility* of identification may occur. As the different practices both within and across countries with regard to what is considered personal and anonymous information to a large extent explain the variation in data access regimes and framework conditions for scientific research across Europe, this is not a trivial issue. On the contrary, this provision could potentially be very important to research, especially in relation to the exchange and sharing of research data across countries.

The various definitions of anonymity very often constitute an obstacle and serious barrier for comparative research projects. One example is the problems involved in using register data in transnational surveys because of various definition of anonymity, another is the different consent requirements depending on whether or not the study fall within or outside the scope of the law.

Consequently the proposal may unintendedly contribute to sustaining and further develop the current fragmentation instead of ensuring identical conditions and practices in line with the legislators' intention.

Sensitive data

In addition to the direct changes in the definition of personal data, a further specification has been added in Article 9(1) regarding special categories of data. Here, the LIBE Regulation elaborates and expands on the definition of personal data. In addition to race/ethnic origin, political opinions, religion or beliefs, it also suggests including *philosophical beliefs*, *sexual orientation or gender identity*, *trade-union membership and activities*, and the processing of *biometric* data. Administrative sanctions, judgments, criminal or suspected offences have also been added to the definition. In fact, the article states that the processing of these special categories of data shall be *prohibited*, while exemptions from the prohibition are listed in Article 9(2a-j). Among these exemptions are those that explicitly refer to research (2h-i): *"…paragraph 1 shall not apply if…processing of data concerning health is necessary for health purposes and subject to the conditions and safeguards referred to in Article 81; or…processing is necessary for historical, statistical or scientific research purposes subject to the conditions and safeguards referred to in Article 83; or…processing is necessary for archive services subject to the conditions and safeguards referred to in Article 83a".*

From a research and data preservation point of view these exemptions are important and valuable. However, some of the "conditions and safeguards" that are mentioned in Article 81 and 83 requires that researchers and data curators install and implement high data protection measures and techniques: "*…personal data may be processed for historical, statistical or scientific research purposes only if: …data enabling the attribution of information to an identified or identifiable data subject is kept separately from the other information under the highest technical standards, and all necessary measures are taken to prevent unwarranted re-identification of the data subjects*" (Article 83(1b)). These measures may have serious impact on several research areas. Issues concerning Articles 81 and 83 are treated separately, see chapter 2.3.3 below.

In a statement[36] to LIBE, the Economic and Social Research Council (ESRC) points out that it is often through the delicate analysis of personal data relating to sexual orientation, political beliefs, and race etc. that "…the prejudice against diversity in society can be eradicated" and that this is often dependent on using the data in an identifiable form and citing court judgments, for example, which contain identifying details such as those listed in article 9(1) but have not been put into the public domain by the data subject themselves. Hence, from the view of ESRC it is of crucial importance that the *exemptions* that are stated in Article 9(2a-j) are protected in a feasible way. If not, the regulation may restrict and hinder research in subjects such as law, contemporary history, sociology and political science.

Pseudonymised data

Several non-commercial research organisations and academics[37] have expressed concerns that the amendments to Articles 4(2), 4(2a) and Recital 23 do not recognize that "…pseudonymised data in research are often used in a very robust system with strict organisational, legal and technological safeguards to protect privacy". And further, that the amendments take an oversimplified view of the use of pseudonymous data in research and does not add clarity. In their view, including such robustly pseudonymous data in the scope of the Regulation will impose a disproportionate regulatory burden on this research. This could undermine sophisticated data sharing infrastructures and research "safe havens" - such as data centres, repositories or archives that provide expertise and support services for processing, preserving and disseminating research data. As such, these stakeholders prefer a more risk proportionate approach that could incentivise more sophisticated pseudonymisation practices to enhance privacy.

The definitions of 'personal data' and 'pseudonymised data' are closely related to the material and territorial scope of the regulation, as it defines and delimits the conceptual border of the Regulation. As the final LIBE Regulation expands its scope compared to the COM proposal (it applies to the processing of personal data, *irrespective of the method of processing*; and *whether the processing takes place in the Union or not*) it is of crucial importance that the definitions are precise and clearly demarcated.

2.3.2 Purpose specification and limitation

Original purpose and compatibility

The principle of 'purpose limitation' in the EU data protection framework is laid out and explained in a report from the Article 29 Data Protection Working Party (WP29)[38]. In it, the concept of purpose limitation is explained as a measure to protect data subjects by setting limits on how data controllers are able to use their data while also offering some degree of flexibility for data controllers. Broadly speaking, the concept of purpose limitation has two main building blocks: personal data must be collected for 'specified, explicit and legitimate' purposes (purpose specification) and not be 'further processed in a way incompatible' with those purposes (compatible use)[39].

---

[36] ESRC: Response to the European Commission's proposed European Data Protection Regulation

[37] Protecting health and scientific research in the Data Protection Regulation - Position of non-commercial research organisations and academics
http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/WTP055584.pdf

[38] The Working Party was set up under Article 29 of Directive 95/46/EC. It is an independent European advisory body on data protection and privacy. Its tasks are described in Article 30 of Directive 95/46/EC and Article 15 of Directive 2002/58/EC.

[39] Article 29 Data Protection Working Party, WP203 (00569/13/EN), *Opinion 03/2013 on purpose limitation:*
http://idpc.gov.mt/dbfile.aspx/Opinion3_2013.pdf

This does not necessarily mean that further processing for a different purpose is incompatible: compatibility needs to be assessed on a case-by-case basis based on the following key factors:

- The relationship between the purposes for which the personal data have been collected and the purposes of further processing;
- The context in which the personal data have been collected and the reasonable expectations of the data subjects as to their further use;
- The nature of the personal data and the impact of the further processing on the data subjects;
- The safeguards adopted by the controller to ensure fair processing and to prevent any undue impact on the data subjects.

In the 95 Directive the purpose limitation principle was explicitly expressed in Article 6(b), and the wording in the legal text ("…*further processing of data for historical, statistical or scientific purposes shall* ***not be considered as incompatible*** *provided that Member States provide appropriate safeguards"*) has been interpreted as providing research purposes a special legitimate status. That is, research purposes have been considered as *always compatible with the original purpose* of the data.

The COM proposal continued this practice through the provision of a broad exception from the requirement of compatibility through Article 6(4) and Recital 40. Here it was stated that the processing of personal data for other purposes (than the initial ones) should be allowed "…*in particular where the processing is necessary for historical, statistical or scientific research purposes*" (see Appendix 1 for full legal text).

However, the WP29 recommended that the proposed broad exception from the requirement of compatibility should be deleted, as it would "…severely restrict its applicability and risk eroding this key principle". In the final LIBE proposal, both Article 6(4) and Recital 40 are deleted.

During the processing of the amendments in the LIBE Committee, other EU Parliament committees expressed opinions with regard to the importance of exemptions for scientific research. The Committee on the Internal Market and Consumer Protection expressed that the "…processing of personal data collected to another purpose can be made available for public scientific research when a scientific relevance of the processing of the collected data can be documented", as long as 'privacy by design' is taken into account in the publication of data[40].

The deletion of Article 6(4) and Recital 40 in combination with amendment 101 (Article 7(4)) in the final LIBE version can be considered as a reduction of the purpose compatibility of further processing for research; it seems that the re-use of research data from various sources no longer are granted a special legitimate status. This may require a new, separate legal basis for any further processing. This, in combination with the amendments to Articles 81 and 83 (see below), may put restrictions on the use and sharing of existing data in a wide variety of research, including researchers and scholars of arts and humanities, information science, linguistics and social studies. It may limit the use and reuse of existing data material for research purposes, like audio-visual content, and all kinds of user-generated data from blogs, social media and text data in general, and it may complicate the fusion and linkage of data, e.g. the linking of survey data with administrative data. It complicates researcher

[40] See OPINION OF THE COMMITTEE ON THE INTERNAL MARKET AND CONSUMER PROTECTION (28.1.2013), Amendment 18, Recital 40 b: http://www.europarl.europa.eu/sides/getDoc.do?type=REPORT&mode=XML&reference=A7-2013-0402&language=EN#title5

access to registry data, as the registry holder in theory has been granted the opportunity to refuse access to data, since the reuse may imply an alteration of the original purpose. It opens for a more restrictive interpretation of the law when it comes to granting access to personal data.

Data from administrative and statistical registries often per se means a deviation from the original purpose. The 95 Directive has been working as an explicit legal reference point that confirms that research always is *compatible with the original purpose,* underlining the general importance of research. When this specification is removed, the special position of research is no longer explicitly visible.

Purpose and consent

In the proposed regulations, it is a requirement that consent must be limited to one or more specified purposes (Article 6(1a), Article 9(2a) and Article 7(4)). What this entails is somewhat unclear, but the legislators' intentions seem to be that one can only consent to one or more specified purposes and that the consent is limited to that or those purpose(s). This might mean that broad consent no longer is considered an acceptable option. If that is the case, it might be unfavourable for SSH research, especially for major population surveys within the social and health sciences[41].

An important ethical issue in SSH and health research is unforeseen possible future use and reuse of research output, when no or limited consent is obtained from the data subject (e.g. for biological samples stored in biobanks and/or the linking of administrative data with survey data, gaining new information). *Informed consent* is ethically important to protect the interests of the data subject, protect the confidentiality of personal information, ensure subject autonomy, define research and social interests in the general advancement of knowledge, and maintain public trust in researchers and institutions[42]. *Broad informed consent* (in its broadest sense) expands informed consent by allowing sample use in unforeseen future studies.

The extended restrictions on broader consent suggested in the LIBE proposal (i.e. the requirement that consent must be limited to one or more specified purposes) may be viewed as a step away from the broad consent model that is formulated in the 95 Directive, towards a more 'dynamic consent' model. The goal is to fulfil the individual's need for autonomy by increasing the user participation and strengthening the ownership of personal data by providing the data subject with information for every re-use of their data.

The idea of dynamic consent is to use modern communication strategies to inform, involve, offer choices, and obtain consent for every research project based on available resources[43]. A more dynamic consent model may complicate important health and social science research, especially when seen in light of the purpose-limitation specified in Article 7. Article 7(4) states that "…*consent shall be purpose-limited and shall lose its validity when the purpose ceases to exist or as soon as the processing of personal data is no longer necessary for carrying out the purpose for which they were originally collected*".

[41] CESSDA Position Statement: EU Parliament vote on new data protection legislation: http://www.cessda.net/news/EU-PrivacyRegulationNegotiatingMandate_Implications.pdf

[42] Petrini, C. (2010): "Broad" consent, exceptions to consent and the question of using biological samples for research purposes different from the initial collection purpose. *Social Science & Medicine*, Volume 70, Issue 2, January 2010, Pages 217–220. DOI: 10.1016/j.socscimed.2009.10.004

[43] Steinsbekk, K. S., Kåre Myskja, B. & Solberg, B. *European Journal of Human Genetics* (2013) **21**, 897–902: "Broad consent versus dynamic consent in biobank research: Is passive participation an ethical problem?" http://dx.doi.org/10.1038/ejhg.2012.282

This may turn out to be an obstacle, especially for large scale research projects in medicine and social sciences, as they traditionally have been dependent on obtaining broader consents from participants. Additionally, it is uncertain how the purpose limitation will affect the ethical review of each specific research project by an independent ethics committee, institutional review board or a data protection agency.

Data storage and archive purposes

A basis for the processing of personal data by archive services has been added in the LIBE proposal, in which storage for research purposes is specifically mentioned (Article 9(2a-j), Article 83(a) and Article 5(e)). It is stated that personal data can be stored for longer periods "*…insofar as the data will be processed solely for historical, statistical or scientific research or for archive purposes*". From the point of view of data archives and repositories, it is positive that separate grounds have been included for the processing of personal data by archive services, in which storage for research purposes is specifically mentioned. This is new and clearly strengthens the legitimacy and the framework conditions for national infrastructure services and research data archives such as e.g. CESSDA and CLARIN member institutions. At the same time it is stated that the archiving exemptions are dependent on, and legal only if "*…appropriate technical and organizational measures are put in place to limit access to the data only for these purposes (storage minimization)*".

The introduction of a requirement for technical and organisational measures to allow indefinite storage, may contribute to increase the threshold for data processing for researchers and archives/repositories. The content of "technical and organizational measures" is not specified; neither is it specified what it would mean to sufficiently comply with the requirement. This legal uncertainty is likely to make research organisations and archives less willing to continue to store data even where it may be useful for research in the future. This may subsequently lead to the loss of valuable data resources. Both the ICO and Science Europe[44] oppose the suggested wording in Article 5(e), on grounds that it has the potential to considerably increase the administrative and regulatory burden for research and data preservation without providing further levels of individual protection in an already highly regulated area.

On the other hand this requirement may be seen as a response and support to the recommendation from the OECD[45]. The principles and guidelines from OECD recommend a division of labour between research and research management and that long-term storage and re-use arrangements are trusted to professional research data archives. These recommendations are similar to the ESFRI roadmap process where support to the implementation of strong and sustainable research infrastructures, including deposit facilities ensuring long-term access to and sharing of research data, are important goals. As such, seen in light of recommendations and recent developments in the European research infrastructure progress, these amendments may actually support the development of stronger national research data archives.

2.3.3 Secondary processing and exemptions for research

Sensitive personal information is subject to special demands to protection. Both the COM and LIBE proposals explicitly state that personal information can only be processed if the subject (the 'registered') gives explicit consent. As stated in Recital 42: *"Derogating from the prohibition on processing sensitive categories of data should also be allowed if done by a law… for health purposes,*

[44] Science Europe: *Position Statement: On the Proposed European General Data Protection Regulation*, May 2013.
[45] OECD: OECD Principles and Guidelines for Access to Research Data from Public Funding: http://www.oecd.org/science/sci-tech/38500813.pdf

including public health and social protection and the management of health-care services,…for historical, statistical and scientific research purposes, or for archive services". The exemption for archival services is new in the final LIBE Regulation. This clarification is important from a research and a research infrastructure perspective, as it underlines the public importance of research and emphasis that research has a legitimate need for processing and preserving personal information. The inclusion of an explicit exemption for the processing of personal data by archive services, in which storage for research purposes is specifically mentioned, also strengthens the legitimacy and the framework conditions for research infrastructures and research data archives.

However, this view may be modified when taking into consideration the conditions for processing of personal information that are listed in Article 83 and the conditions for processing of health information that are listed in Article 81.

Article 81(2) narrows the exemption from consent for the use of data concerning health in research, stating that "…*processing of personal data concerning health which is necessary for historical, statistical or scientific research purposes shall be permitted only with the consent of the data subject*". However, exemptions from the requirement for consent can be granted for research that serves a high public interest (Recital 123(a)). As stated in Article 81(2a) "…*Member States law may provide for exceptions to the requirement of consent for research, […] with regard to research that serves a high public interest, if that research cannot possibly be carried out otherwise".* Although these exemptions are valuable and decisive from a health research point of view, the notion that national law may provide for exemptions to the requirement of consent is somewhat hard to interpret in practical terms as it seems to contradict the initial purpose of the new Regulation, namely the harmonisation of the European legal system for data protection.

It should be noted that an anonymisation or pseudonymisation requirement applies, regardless of whether the processing is based on consent, which is the main rule, or the processing is exempted from consent.

The requirement for pseudonymisation to be at "the highest technical standards" may be problematic because research resources are often used over many years. Even where a study can demonstrate "highest technical standards" when it is first established "…it would be impractical to ensure compliance with this requirement every single time data are used in the future, as this would require continual updating of standards and processes[46]."

Although exemptions from the requirement for consent can be granted for research that serves a "high public interest" (as stated in 81(2) and Recital 123), it also suggests that the exemption is to be used in a(n) (unspecified) limited set of circumstances only. This may be problematic for several studies, especially in medicine and health related research as the results and impact of the study are not known at the outset.

In addition, Article 81(1b) is new in the LIBE proposal and introduces further special consent rules for the use of data concerning health. The wording "…*where the data subject's consent is required for the processing of medical data exclusively for public health purposes of scientific research, the consent may be given for one or more specific and similar researches*" is somewhat ambiguous and

[46] Protecting health and scientific research in the Data Protection Regulation - Position of non-commercial research organisations and academics:
http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/WTP055584.pdf

could possibly create a lack of clarity around whether the consent used for a particular study would comply to the reuse of data in other similar studies.

Both the Committee on Internal Market and Consumer Protection[47] and the Committee on Legal Affairs[48] wanted to include a reference to Articles 81 and 83 in Article 5(e), on grounds that it should be possible to store personal data for longer periods *for health purposes* (Article 81) as well as for historical, statistical and scientific research purposes (Article 83). In the MCP Committee's justification it was expressed that "this will ensure that all relevant data is available to deliver the most appropriate care to the data subject"[49].

2.4 GDPR and preservation policies: conclusion

Although there is still some uncertainty connected to the further processing of the Regulation, the European Council aims for its adoption in late 2014 and the regulation is planned to take effect after a transition period of approximately two years. Hence, it is expected to be implemented and takes effect in 2016.

As we have seen, in the final LIBE regulation the provision specifying that research shall not be considered as incompatible has been removed from the regulation text altogether. Seen in conjunction with other amendments in the final regulation, particularly those that tighten requirements for consent for the processing of health data and sensitive data, it seems that the balance in the legislation has shifted in favour of data protection at the expense of research opportunities.

Another concern is the fact that the LIBE legislation leaves an opening for individual Member States to adopt exemption provisions that safeguard research needs. This somewhat contradicts the original intention of the Regulation which was to increase the harmonisation in the current fragmented data protection area. The likelihood that these differences will continue to exist and that differences in general conditions for research will be maintained are increased within the LIBE regulation. This may have an impact on the research sector's opportunities to contribute to knowledge development both at the national level and in Europe.

[47] See OPINION OF THE COMMITTEE ON THE INTERNAL MARKET AND CONSUMER PROTECTION (28.1.2013), Amendment 67, Article 5 e: http://www.europarl.europa.eu/sides/getDoc.do?type=REPORT&mode=XML&reference=A7-2013-0402&language=EN#title5

[48] See OPINION OF THE COMMITTEE ON LEGAL AFFAIRS (25.3.2013), Amendment 46, Article 5 e: http://www.europarl.europa.eu/sides/getDoc.do?type=REPORT&mode=XML&reference=A7-2013-0402&language=EN#title6

[49] Ibid., Amendment 68, Article 6(c).

3 Copyright issues in preservation policies

3.1. Introduction

The collection, archiving, dissemination and re-use of research data may face several issues related to copyright. Depending on national legislations and the work agreement between researchers and their employers, the producers of research data or their organizations can claim copyright to original data. In many cases, researchers in the social sciences and humanities will let other researchers get access to their data, either by putting the data in the public domain, or by retaining copyright but licensing the data, often without financial compensation, but sometimes with an embargo period. When data are archived in repositories, these repositories will need to take copyright and licensing conditions into account.

In 2007, the Organisation for Economic Co-operation and Development (OECD) published the "OECD Principles and Guidelines for Access to Research Data from Public Funding" with the aim of establishing international guidelines on access to research data.[50] These guidelines state:

> "*The nature of 'public funding' of research varies significantly from one country to the next, as do existing data access policies and practices at the national, disciplinary and institutional levels. These differences call for a flexible approach in developing data access arrangements. The balance between the costs of improved access to research data and the benefits that result from such access will need to be judged by individual national governments and their communities.*"

The OECD guidelines also state how the protection of intellectual property should be carried out:

> "*Data access arrangements should consider the applicability of copyright or of other intellectual property laws that may be relevant to publicly funded research databases. Factors to consider include:*
>
> - *As public/private partnerships in the funding of research and related data production are increasing, balanced public/private arrangements should facilitate broad access to research data where appropriate. The fact that there is private sector involvement in the data collection should not, in itself, be used as a reason to restrict access to the data. Consideration should be given to measures that promote non-commercial access and use while protecting commercial interests, such as delayed or partial release of such data, or the voluntary adoption of licensing mechanisms. Such measures can allow the primary participants to fully exploit the research data without unnecessarily shutting off access.*
> - *In those jurisdictions in which government research data and information are protected by intellectual property rights, the holders of these rights should nevertheless facilitate access to such data particularly for public research or other public-interest purposes.*"

The Community Research and Development Information Service (CORDIS) of the European Commission also has *a Guide to Intellectual Property Rules for FP7 projects*. It is "*a guide to the*

[50] OECD (2007). OECD Principles and Guidelines for Access to Research Data from Public Funding. Retrieved 23 June 2014, from: http://www.oecd.org/science/sci-tech/38500813.pdf

various issues and potential pitfalls regarding IPR that participants may encounter when preparing and participating in an FP7 project".[51]

However, despite general guidelines on IPR issues that are at the disposal of researchers, these issues are not always easy to tackle and solve. The situation is often more complicated in the case of annotated language data, which is a rather common type of research data in projects like CLARIN, one of the five ESFRI actions in DASISH. Annotated language data consist of text, spoken, or multimodal corpora, which, for study purposes, are enriched by means of transcription, translation, alignment, part-of-speech tagging, parsing or other linguistic analysis. Although annotators may claim ownership of the annotations, the original source texts are usually authentic works of fiction or non-fiction, copyrighted by their authors or publishers. Thus, it may be necessary to obtain permission to effectively study certain works with digital means, and to preserve the results of the research carried out on those copyrighted works.

Whether and under which conditions researchers may be entitled to make copies of copyrighted works for research purposes depends on the country, as will be illustrated in section 2. Current efforts to harmonize and reform copyright legislation in Europe will be sketched in section 3. An overview of licensing schemes as practiced in current research infrastructures and repositories will be given in section 4.

3.2. Copyright exceptions in different countries

Most countries are signatories of the International Convention for the Protection of Literary and Artistic Works, commonly known as the *Berne Convention,* an international copyright agreement under which all contracting countries provide protection to those works published in the signatory countries, as well as to unpublished works by citizens of residents of such countries.[52] However, the extent of copyright as well as possible exceptions, often under the headings 'fair use' or 'fair dealing', depend on national legislation. Although a full overview is beyond the scope of this document, relevant legislation in a few countries will be briefly compared below.

It must also be taken into account that copyright may have different expiration dates in different countries, such that works may enter into the public domain at different times. Several factors determine whether a work may be considered to be in the public domain, and sometimes the expiration dates also depend on the nature of the work.[53] For instance, sometimes a distinction between written texts or videos is established. It may also be the case that a work has entered the public domain, but has newly been released in some other form (e.g. a critical edition of a classic work, or a new version of a piece of music), which subsequently has acquired some Intellectual Property Rights (IPR) which must be complied with. Therefore, prior to using any work, the national legislation of the country where such work originates has to be consulted. EU Law, for instance, establishes that copyright expires 70 years after the death of the copyright owner.

[51] CORDIS. Guide to Intellectual Property Rules for FP7 projects. Retrieved 23 June 2014, from: http://ec.europa.eu/research/participants/data/ref/fp7/89593/ipr_en.pdf

[52] World Intellectual Property Organization. Berne Convention for the Protection of Literary and Artistic Works. Retrieved 23 June 2014, from: http://www.wipo.int/treaties/en/text.jsp?file_id=283698.

[53] Mannapperuma, Menesha A., Schofield, Brianna L., Yankovsky, Andrea K., Bailey, Lila and Urban, Jennifer M. (2014). Is it in the Public Domain? A handbook for evaluating the copyright status of a work created in the United States between January 1, 1923 and December 31, 1977. Berkeley Law, University of California and Samuelson Law, Technology and Public Policy Clinic. Retrieved 23 June 2014, from: http://www.law.berkeley.edu/files/Final_PublicDomain_Handbook.pdf

In order to use copyrighted works (copying, adapting, sharing, processing, etc.), the permission of the copyright owner shall be obtained. This may prove to be challenging, especially when dealing with complex copyright works.

EU law also establishes some exceptions for re-use of copyrighted works that *can be implemented* within Member States (i.e. the laws are country-dependent). Such exceptions are:

- Criticism and review
- News reporting
- Private copying
- Parody
- Research
- Education
- Archiving and preservation

As copyright is currently country-dependent under EU copyright law, the laws of all countries of origin of the different sources of data are to be respected. Thus, depending on the country of origin, different uses may be allowed and various restrictions may apply.

3.2.1 USA

The USA has a judicially established practice of "*fair use*" that makes it legal to make copies of copyrighted works without explicitly obtaining permission from the rights holders.

The importance of a *fair use* provision for research was shown in a lawsuit by the Authors Guild et al. v. Google on the case of Google Books. In November 2013, judge Denny Chin ruled in favour of Google Inc. as he deemed that Google scanning books and digitizing them was allowable since it permits scholars to analyse massive amounts of data, while not infringing the rights of the authors of such texts. An investigation of n-gram frequencies as empirical material for humanities studies was explicitly mentioned as an example of fair use.[54]

On June 10, 2014, the U.S. appeals court ruled against authors in another book-scanning lawsuit known as the HathiTrust case. The HathiTrust Digital Library consists of 80 member institutions, which made the books in their collections available for inclusion in the digital library. This digital library offered three main uses: full-text search, access to the print-disabled, and preservation. According to the appeals court, HathiTrust did not violate copyright protections and their scanning activities were rather a fair use of copyright works.[55]

3.2.2 UK

According to the laws of most countries, including Britain, the permission of the copyright owner shall be obtained in order to use (copying, adapting, sharing, processing…) copyrighted works. This may prove challenging, especially when dealing with complex copyrighted works. When the copyright owner is either unknown or cannot be located, such works may be referred to as *orphan works*. Orphan works cannot be re-used unless the use is covered by an existing exception to copyright.

The United Kingdom has partially implemented some of the exceptions that EU law allows its member states to implement. On 1 June, 2014, several new regulations came into force. These

[54] United States District Court Southern District of New York (2013). Authors Guild et al. v. Google. Lawsuit on the case of Google Books. November 2013. Retrieved 23 June 2014, from http://www.wired.com/images_blogs/threatlevel/2013/11/chindecision.pdf

[55] United States Court of Appeals for the Second Circuit (2014). Authors Guild v. HathiTrust Appeal Decision. 11 June 2014. Retrieved 23 June 2014, from: https://www.eff.org/files/2014/06/10/agvhathitrust.pdf

Regulations include provisions as regards disability; research, education, libraries and archives; and public administration.[56]

As far as research is concerned, the new regulations in UK have extended the scope of its exceptions and now all copyright works, including sound recordings, films and broadcasts are included. Moreover, the scope is also broadened to include copies provided to users by libraries. The new provision establishes that these works can be used for research, provided this usage can be considered "fair". "Fair dealing" is defined in the new regulations as the usage a fair-minded and honest person would do with the work.

The new copyright exception permits UK researchers to carry out non-commercial research using *text and data mining* technologies. This new provision establishes that whenever a person has lawful access to any copyrighted material, no additional permission from the rights holders will be needed to copy the work for text and data mining for non-commercial research as long as the use of the aforementioned work is sufficiently acknowledged.

3.2.3 Norway

In Norway, the copyright of data expires 70 years after the death of its author. Thus, in order to re-use any kind of copyrighted data, it is required the existence of a license or a statement of what is and is not allowed to do with such data. However, it is possible to ask the Government for a research exemption granting access to copyrighted materials. The regulations to the Copyright Act, §1-4, establishes that the Ministry of Culture can grant research institutions the right to access and use copyrighted texts for research purposes:[57]

> *[NO] Forskrift § 1-4 "Kultur- og kirkedepartementet kan for forskningsformål gi undervisnings- og forskningsinstitusjoner tillatelse til å fremstille eksemplar av åndsverk, også i andre format enn originaleksemplaret."*
>
> *[EN] Regulations § 1-4 "The Ministry of Culture and Church Affairs may, for research purposes, grant education and research institutions permission to copy copyrighted works, also in formats different from the original."*

The University of Oslo has been granted such an exemption twice, e.g. in the case of the Oslo Parallel Corpus.

3.3. European reform

Europe does not have a unified legal framework for copyright, but there are different legal frameworks in the various European countries. These differences create barriers for cooperative R&D across borders within Europe, in particular for text and data mining. Digital materials and services pose new challenges to the interpretation of copyright legislation while Europe has no provision for *fair use* of materials protected by copyright.

On December 5, 2013, the European Commission launched a public consultation as part of its on-going efforts to review and modernise the EU copyright rules. A consultation document with questions was made available and all stakeholders were welcome to contribute to this consultation, which closed on March 5, 2014. The number of responses was 9599.

---

[56] Intellectual Property Office (UK) (2014). Changes to copyright law and guidance. Retrieved 23 June 2014, from: http://www.ipo.gov.uk/copyright-exceptions.htm

[57] http://lovdata.no/dokument/SF/forskrift/2001-12-21-1563/

CLARIN ERIC, one of the ESFRI actions involved in DASISH, contributed a response to this consultation. So did several national CLARIN consortia, including, for instance, CLARINO, the project for the construction of CLARIN in Norway. In their replies, the importance of a reform for the research in social sciences and humanities is stated, and the need for a EU-wide *exception* allowing copyrighted materials to be used in research is voiced and argued.

As pointed out by CLARIN's and CLARINO's reply to the consultation, holders of copyrights and related rights do not enjoy a single protection in the EU and are protected on the basis of a bundle of national rights in each Member State. Thus, rights are to be acquired and enforced on a country-by-country basis under national law. This situation hampers researchers and institutions aiming at offering European-wide services, as rights have to be cleared in several countries prior to making such services available in each country. Currently, it might be the case that the same service cannot be accessed from different countries. The main challenge is thus to "*increase the cross-border availability of content services in the Single Market, while ensuring an adequate level of protection for right holders*."

A related issue voiced by CLARIN and CLARINO is the importance of access to *big data*. While some other countries like the United States, the United Kingdom and Australia have a *fair use* provision, most EU member states do not have such a provision, but require researchers and educators to make license agreements with all rights holders. Since the latter requirement is inoperable for big data, countries allowing simpler access to big data under the *fair use* provision will promote the dominance of their research and innovation. In practice, this will continue to favour research on the English language.

CLARIN states that access to big data is crucial for research, but unfortunately "the reality is that the material available is much too small. A vast body of material which is interesting from a research perspective is language material from the 1860's and onward which is out of print but whose copyright may not yet have expired and whose owners may or may not be traceable at costs which are out of proportion to the business interests the copyright is intended to protect. In addition, most of the modern language material from the 1980's and onward, which can be licensed, with some effort is tied to individual research groups and institutions in the Member States. In order to provide the same material to other researchers with similar non-commercial research interests EU-wide, licensing agreements have to be renegotiated. In addition, copyright is intended to protect business interests, but non-commercial research is by definition non-commercial, so no commercial interests are at stake and copyright licensing is not even an appropriate vehicle for furthering this to begin with."

The same is voiced by CLARINO: "The current problem is that most of the licensed language material is tied to individual research groups and institutions in the Member States. In order to provide the material EU-wide, the licensing agreements have to be renegotiated, which is often very expensive and time and labour consuming. Even if the new agreements we are negotiating with some right holders aim to solve a small part of the problem, the vast body of previously licensed language material remains a problem."

For the reasons stated above, both CLARIN and CLARINO argue that the best way forward is the establishment of a mandatory exception for research use. This argument, which has earlier been

voiced in the Hargreaves report[58] preparing changes in the UK, is also voiced by the European Copyright Society in their answer to the consultation: "... it may well be that the only way to overcome both problems (non-harmonized and fragmented national solutions and territoriality/choice of law) and ensure a real internal market for teaching and research services – at least, online – would be by means of a **mandatory and uniform exception / limitation for teaching and research purposes within the EU territory**."[59]

3.4. Licensing schemes

To the extent that research data contain materials, which are not in the public domain and no lawful exception applies, permission needs to be obtained to copy, preserve and distribute the data. For data deposited in repositories, various parties are usually involved, so that in general, two main agreements are required:

1. A deposition license agreement (DELA) is an agreement between the *repository* and the *owners* of the rights to a deposited resource. It regulates the conditions under which the resource will be made available. As part of these conditions, it specifies which EULA (see below) will be applied.
2. An end user license agreement (EULA) is an agreement between the *repository* and the end *users* of a deposited resource. It regulates the conditions under which the users can access and exploit the resource. Terms of use which are not specific to a particular resource, but which apply for all repository services, are usually brought together in the terms of a service agreement (TOS).

By *license*, in this context, one usually means the conditions set forth in the EULA. Deliverable D 4.5.1 of the QT LaunchPad project[60] offers a good overview of the types of licenses currently available. The widely known and used Creative Commons licenses are standard licenses which cannot be modified. They enable quick and easy licensing of resources. In contrast, there also exist license templates which can be modified according to specific needs that arise in negotiations. The Meta-Share and the CLARIN templates, discussed below, are examples of these.

3.4.1 Creative Commons

Creative Commons are standard licenses. They are based on four non-mandatory core elements[61]:

- **Attribution**: This license lets others distribute, remix, tweak, and build upon the work, even commercially, as long as they credit the licensor for the original creation.
- **Non-commercial**: This license lets others remix, tweak, and build upon the work non-commercially.
- **Share-Alike**: This license lets others remix, tweak, and build upon the work, as long as they license their new creations under identical terms.

[58] Hargreaves, Ian (2011). Digital Opportunity: A Review of Intellectual Property and Growth. Retrieved 23 June 2014 from http://www.ipo.gov.uk/ipreview-finalreport.pdf

[59] ECS (2014). European Copyright Society Answer to the EC Consultation on the review of the EU copyright rules. Retrieved 23 June 2014, from: http://www.ivir.nl/nieuws/ECS_EC_consultation_copyright.pdf

[60] Tsiavos, Prodromos; Piperidis, Stelios; Gavrilidou, Maria; Labropoulou, Penny; and Patrikakos, Tasos (2013). Deliverable D 4.5.1: Legal Framework. Retrieved 23 June 2014, from: http://www.qt21.eu/launchpad/system/files/deliverables/QTLP-Deliverable-4_5_1_0.pdf (pp. 21–27)

[61] For more information about the Creative Commons Licenses and the combination of the different elements see: http://creativecommons.org/licenses/

- **No derivatives**: This license allows for redistribution, commercial and non-commercial, as long as it is passed along unchanged and in whole.

The licensor may choose whichever combination of these elements he/she deems best for his/her resource.

3.4.2 Meta-Share

As it is stated in their website, the META-SHARE licensing scheme[62] consists of three different sets of licenses depending on the resource end-users. The first set is based on the Creative Commons licenses and is aimed at any type of end-users, without restricting them to a particular community. The second set of licenses aims at allowing META-SHARE members and resource depositors to make their resources available to other network members only. It consists of a set of META-SHARE Commons Licenses based on the first layer of Creative Commons licenses. The third set of licenses consists of a set of no redistribution licenses that allow use and exploitation of the resources and at the same time ensure that the resource owner has full control over the resource distribution.

In addition to the three different sets of licenses within METASHARE, as stated in the META-SHARE website, "*a set of legal document templates (non-licences) is offered that is designed to help all stakeholders (resource owners, distributors and end-users) work in a friendly and transparent environment. These include a Depositor's Agreement (DA), a Memorandum of Understanding (MoU) for the Network members and a Service Level Agreement (SLA)*."

Further, the META-SHARE also had an IPR Helpdesk to assist researchers in understanding, choosing and using any of the licenses and other legal tools. Finally, and also as stated in the META-SHARE Memorandum of Understanding, "*resources should ideally be open or shared at least for research purposes. The copyright conditions of the initial raw resource should be known, and in any case the Depositor should have all rights that allow for distribution through a network like META-SHARE. Likewise, processed and derivative resources (annotated web or other text, lexica extracted from parallel text, etc.) should ideally be open at least for academic/research purposes*."

3.4.3 CLARIN

CLARIN has defined its own set of license templates but also allows resources to be deposited under other licenses. The CLARIN license templates have a simple classification in order to make them easily understandable. Materials can be Publicly available (PUB), for Academic use only (ACA) or for Restricted use (RES). Additional conditions may include, for instance, attribution, non-commercial use only, share-alike, the obligation to redeposit within CLARIN, the obligation to inform the resource developer whenever the resource is used and for which purposes, restrictions on downloading, etc.

CLARIN offers the following set of *templates* for agreements:

- The CLARIN Deposition License Agreements (DELA), which are to be signed between the third party and a CLARIN center;
- The CLARIN End-User License Agreements (EULA), which are to be signed between the end user and the CLARIN center; and
- The CLARIN Terms of Service (ToS), which are to be signed between the end user and the CLARIN center.

[62] For detailed information about the META-SHARE licensing scheme see: http://www.meta-net.eu/meta-share/licenses

3.4.4 ESS

While mainly dealing with open data, ESS also establishes some restrictions for potential end-users[63]. Concretely, they indicate that "*the data are available without restrictions, for not-for-profit purposes*". Moreover, they also indicate that all data available is anonymous to comply with the data protection regulations.

3.4.5 SHARE

SHARE also has established several data access rules[64]. Prior to accessing data, applicants must prove that they have a scientific affiliation and have to sign a statement where they commit to use the data only for purely scientific purposes. Additionally, data users are not allowed to make copies, nor to redistribute the data to third parties. There is an obligation to include a disclaimer and an acknowledgement in publications using SHARE data, and the SHARE coordination team must be informed about such publications.

3.4.6 CESSDA

In its statutes, CESSDA[65] declares that data access shall be done compliant to the recommendations and guidelines on data access of the OECD. The data will be available for authenticated members. The Intellectual Property holder of the data will remain the original Intellectual Property holder, and when the data have been originated from CESSDA-funded work, CESSDA AS will be the IPR holder unless otherwise previously agreed upon.

There is currently no CESSDA template or standard for licenses, and as pointed out in the "Access Policies and Licensing for Archives and Repositories" presentation in the DASISH IASSIST Workshop organized in June 2013, the different CESSDA Archives have different licenses as regards re-use of data.

3.4.7 DARIAH

In the case of The Language Archive (TLA) it is also established that IPR has to be duly respected and taken into consideration[66]. To ensure that both moral and legal rights are properly handled, TLA has four different levels of access for its users:

1. Open access, for resources which can be immediately accessed.
2. Restricted access for registered users, for resources which can be accessed by registered TLA users upon agreement of a Code of Conduct.
3. Access upon authorized request, for resources which are protected and for which access has to be granted upon request. The access may be in turn restricted for a specific use or during a limited amount of time.
4. Depositor's access only, for those resources which can only be accessed by their depositors.

3.4.8 Acceptance of End User License Agreements

An essential component for enforcing end user license agreements is the user authentication and authorization process prior to accessing data. A proper authentication and authorization infrastructure (AAI) ensures that only users with the appropriate credentials get access to copyrighted data. Users are also required to read and accept license agreements where all provisions

[63] See http://www.europeansocialsurvey.org/data/conditions_of_use.html for further details.

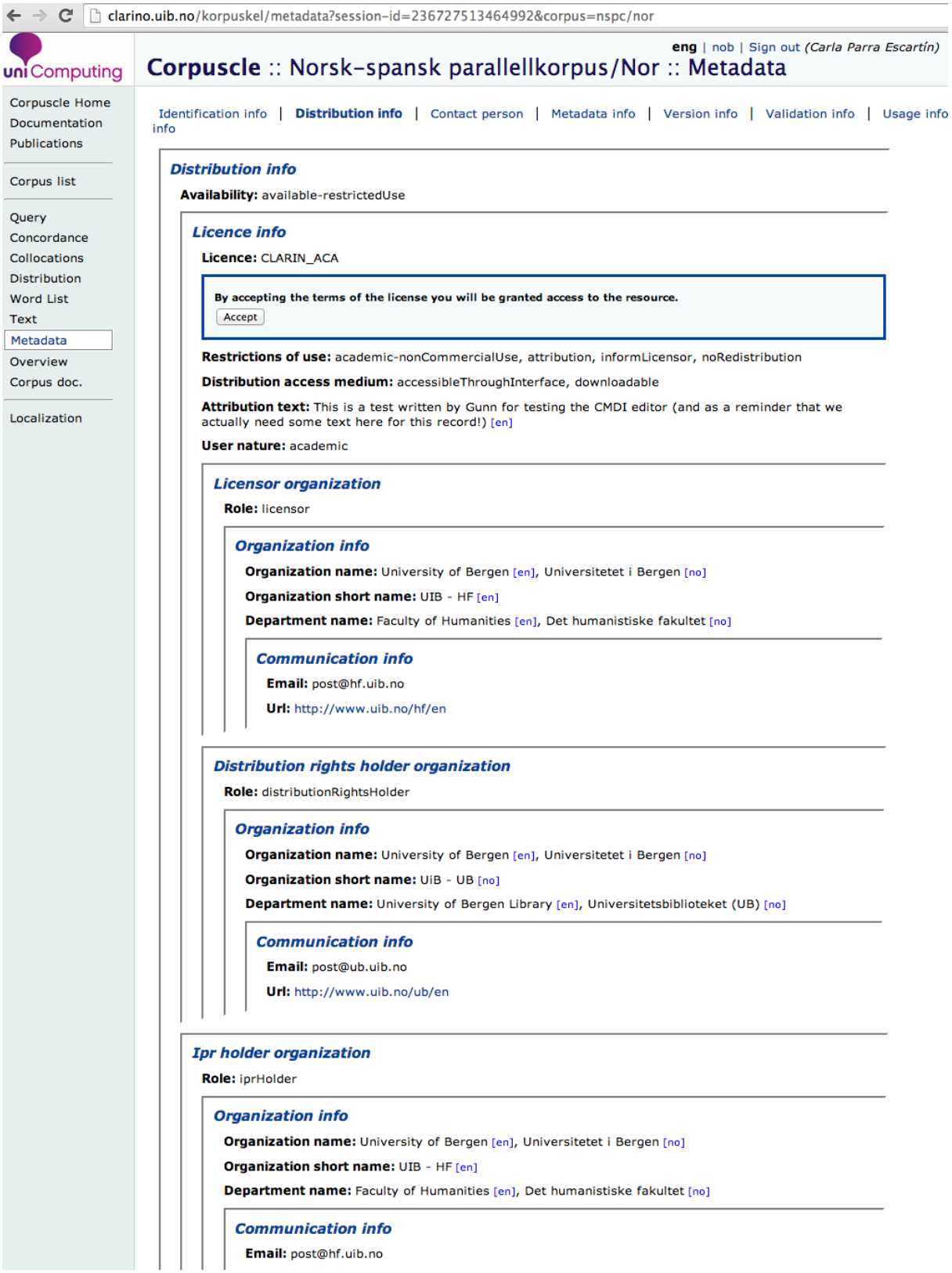
[64] See http://www.share-project.org/data-access-documentation/research-data-center-data-access.html for further details.

[65] See http://cessda.net/about/docs/Statutes-for-CESSDA-18-June-2013-final-version-1.pdf for further details.

[66] See http://tla.mpi.nl/resources/access-permissions/ for further details.

regulating the usage of such data are specified. By accepting the terms of the licenses, users commit to abide by those terms. Click-through licenses require a high level of trust based on a secure AAI (see also the DASISH training module on AAI).

The screenshot below is taken from Corpuscle, a corpus search and visualization tool in CLARINO (the Norwegian CLARIN project); it shows how the terms of a license are presented to the user and can be accepted by clicking.



3.5. IPR Issues in preservation policies: conclusion

For the humanities, the primary IPR bottleneck is copyright. The lack of a uniform European copyright law including exceptions for research poses a challenge for researchers working with copyrighted data. Copyright clearing can be a very cumbersome task, not only because many rights holders may be involved but also because so many different laws have to be taken into consideration.

In the current situation, it is a necessity for infrastructures to establish and implement policies that provide guidance to repositories and their users. CLARIN develops policies through its Legal Issues Committee (CLIC), which is in charge of advising the Board of Directors on all issues related to IPR, privacy protection and ethical matters. Among the tasks of CLIC are to collect and publish information and recommendations related to legal issues, as well as the legal policies adopted by CLARIN. Country-specific information and recommendations on legal issues and policies is also provided by the different national CLARIN projects[67]. Several national helpdesks have also been established. Furthermore, CLIC represents CLARIN as an observer in the World Intellectual Property Organization (WIPO).

Research infrastructures are conscious of the challenges posed by copyrighted data, and are therefore establishing and implementing policies to help and guide researchers who not necessarily are legal experts.

[67] See for instance the German CLARIN project's website (http://de.clarin.eu/en/training-helpdesk/legal-helpdesk.html), or its Finnish counterpart https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/KielipankkiLicenceCategories. Both include information on legal issues.

4 Final conclusions and recommendations

Data archives, data centres and data repositories are currently expanding their role as research data infrastructures and support services for several stakeholders in the data life-cycle, e.g. those creating data and those accessing and consuming data. Within the social science community the data archives have been providing a wide range of services for many years, services needed to ensure easy access to high quality data. In addition to the development of various data access tools, data and metadata standards, trust certifications (e.g. the Data Seal of Approval) and providing training in methodological and analytical techniques, they are providing assistance and advice regarding data management, data depositing, and legal and ethical issues.

This contributes to ensuring the long-term preservation, accessibility and quality of research data. In this process of expanding their roles several of these services have built, and keep building, a strong and varied 'repository' of expertise through experience and direct contact with data and legal/ethical issues. Through a diversity of projects they encounter more data than most researchers or short-term and temporary service providers are likely to ever encounter. Hence the archives and data repositories can be valuable data facilitators and provide support services in several areas, e.g. provide information about various legal and procedural access requirements and support services; provide information and guidance about legal and ethical requirements and practices; provide active help with specific queries; and through proactive help like providing training and support materials. Additionally, they can provide the technological capacity to share data that otherwise would be lost or difficult to share, providing legally binding user licenses and licence agreements that can secure data services for rich but sensitive data.

The best way for data repositories and archives to prepare for the possible practical effects of the GDPR and the current IPR reform is through solid methodologies for enforcing the desired attributes and properties of their collections, i.e. through a policy-based data environment that can build trustworthy collections.

It is important to recognise that a preservation policy is essential for the archive or repository regardless of choices made for how the services are implemented and delivered. Even if distributed service options are used, a digital preservation policy is necessary to frame the requirements for service level agreements and licensing agreements for all levels and involved partners in the distributed preservation service network[68].

The curation, processing and preservation of sensitive data, and clauses concerning intellectual property rights and restrictions on use of repository content, should be clearly stated in well-formulated preservation policies. The policies should be regulated and implemented through deposit agreements, contracts, and licenses. This is necessary in order to allow the repository to track, act on, and verify rights and restrictions related to the use of the digital objects within the repository.

The preservation policy should be more than a general management statement. It should contain general policy clauses and a clear description of how these clauses are to be implemented. There

[68] See, CCSDS 652.0-M-1: Audit and Certification of Trustworthy Digital Repositories. Magenta Book, September 2011. http://public.ccsds.org/publications/archive/652x0m1.pdf. Also an ISO (ISO 16363:2012): http://www.iso.org/iso/catalogue_detail.htm?csnumber=56510

should be a close connection between repository purposes and properties on the one hand, and policies and implementable procedures on the other.

Thus, a preservation policy should define and specify the repository's requirements and processes for managing personal and sensitive data, intellectual property rights, depositor agreement and access and license agreements. In light of the proposed changes to the GDPR and the ongoing European reform on copyright issues and licensing schemes, special considerations should be taken with regard to an explicit and concise specification of the technical and organizational measures that are in place to regulate access to data that are limited to specific purposes (referred to earlier as 'storage minimisation').

There should be a continual updating of standards and processes, and development of the necessary skill and expertise. Staff training and financial and organisational planning for the archive or repository should be clearly stated and should include provision for activities like staff training, technical infrastructure, preservation activities, storage and media (formats) routines, and changes due to evolving technology and legal framework. The preservation policy should also clearly state the legal and regulatory framework(s) under which the repository or archive operates.

Further the policy should include an assertion of copyright and intellectual property rights, and agreements with authors and data owners should be made clear and recorded through explicit agreements with authors on rights for preservation and reproduction of the data. This should be combined with explanations of access levels and access restrictions, and procedures for how different levels are assigned to different datasets or data collections. A commitment to keep data secure should be stated and any changes to data and/or metadata should be tracked.

As established in DASISH D4.1[69] there are (at least) five methods or key standards which digital repositories can use to assess themselves and to support public statements about their level of trustworthiness, ranging from OAIS core conformance to initial self-assessment (DRAMBORA, PLATTER, DSA) to formal audit and certification by external auditors (TRAC, DIN 31644, ISO 16363). These tools and standards, combined with an attentive focus on legal reform and progress, should constitute the core element of the development and maintenance of data preservation policies.

The development of a new GDPR and the ongoing IPR reform, and recent recommendations (e.g. OECD) and developments (e.g. the transferring of assets and activities of the ESFRIs to becoming ERICs) in the European research infrastructure network, seem to imply a professionalization of the SSH data preservation community. The purpose of a professional data deposit and preservation service for scientific use is that the services can comply with requirements for safety, security, longevity and continued access. Services must hence operate in a perspective that spans decades. At the same time research communities as well as the general community need to be assured that data delivered today can be retrieved and used tomorrow - while maintaining the interest of the data subject through solid data protection measures. In order to achieve this objective requires facilities which are well institutionally embedded and can demonstrate a high degree of permanence.

[69] DASISH Deliverable Report D4.1: Roadmap for Preservation and Curation in the SSH

Appendix 1 – Comparison of legal texts

The following is a tabulated comparison of legal texts of the original Commission proposal from 2012 and the final LIBE-approved version of 2014. For illustrative reasons we've included, where possible, the corresponding legal text of the 95 Directive. The legal texts are organized based on the issues discussed in part 1 of this report.

Bold and emphasised text illustrates the differences between the legal proposals (not for the 95 Directive).

	Legal Proposal		
Issue	**95 Directive**	**COM**	**LIBE**
Material scope	Article 3 1. This Directive shall apply to the processing of personal data wholly or partly by automatic means, and to the processing otherwise than by automatic means of personal data which form part of a filing system or are intended to form part of a filing system.	Article 2 1. This Regulation applies to the processing of personal data wholly or partly by automated means, and to the processing other than by automated means of personal data which form part of a filing system or are intended to form part of a filing system.	Article 2 1. This Regulation applies to the processing of personal data wholly or partly by automated means, ***irrespective of the method of processing,*** and to the processing other than by automated means of personal data which form part of a filing system or are intended to form part of a filing system.
Territorial scope	Article 4 1. Each Member State shall apply the national provisions it adopts pursuant to this Directive to the processing of personal data…	Article 3 1. This Regulation applies to the processing of personal data in the context of the activities of an establishment of a controller or a processor in the Union.	Article 3 1. This Regulation applies to the processing of personal data in the context of the activities of an establishment of a controller or a processor in the Union***, whether the processing takes place in the Union or not***.
Definition of personal and pseudonymous data	Article 2 (a) 'personal data' shall mean any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more	Article 4 (1) **'data subject' means an identified natural person or a natural person who can be identified, directly or indirectly, by means reasonably likely to be used by the controller or by any other natural or legal person, in particular by reference to an identification number, location data, online**	Article 4 (2) 'personal data' means any information relating to **an identified or identifiable natural person** ('data subject'); **an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location**

	factors specific to his physical, physiological, mental, economic, cultural or social identity;	**identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that person;** (2) 'personal data' means any information relating to **a** data subject;	**data, unique identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social or gender identity of that person**; **(2a) 'pseudonymous data' means personal data that cannot be attributed to a specific data subject without the use of additional information, as long as such additional information is kept separately and subject to technical and organisational measures to ensure non-attribution;**
	Recital 26 Whereas the principles of protection must apply to any information concerning an identified or identifiable person; whereas, to determine whether a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person; whereas the principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable; whereas codes of conduct within the meaning of Article 27 may be a useful instrument for providing guidance as to the ways in which data may be rendered anonymous and retained in a form in which identification of the data subject is no longer possible;	Recital 23 The principles of protection should apply to any information concerning an identified or identifiable person. To determine whether a person is identifiable, account should be taken of all the means **likely reasonably** to be used either by the controller or by any other person to identify the individual. The principles of data protection should not apply to data **rendered anonymous in such a way that the data subject is no longer identifiable**.	*Recital 23* The principles of **data** protection should apply to any information concerning an identified or identifiable **natural** person. To determine whether a person is identifiable, account should be taken of all the means **reasonably likely** to be used either by the controller or by any other person to identify **or single out** the individual **directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the individual, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration both available technology at the time of the processing and technological development**. The principles of data protection should **therefore** not apply to **anonymous** data, **which is information that does not relate to an identified or identifiable natural person. This Regulation does therefore not concern the processing of such anonymous data, including for statistical and research purposes**.

Definition of sensitive data	**Article 8** **The processing of special categories of data** **1. Member States shall prohibit the processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, and the processing of data concerning health or sex life.**	**Article 9** *Processing of* **special categories of** *personal* **data** **1. The processing of personal data, revealing race or ethnic origin, political opinions, religion or beliefs, trade-union membership, and the processing of genetic data or data concerning health or sex life or criminal convictions or related security measures shall be prohibited** **2. Paragraph 1 shall not apply** *where*: **…** **(h) processing of data concerning health is necessary for health purposes and subject to the conditions and safeguards referred to in Article 81; or** **(i) processing is necessary for historical, statistical or scientific research purposes subject to the conditions and safeguards referred to in Article 83;**	***Article 9*** **Special categories of data** **1. The processing of personal data, revealing race or ethnic origin, political opinions, religion or** *philosophical* **beliefs,** *sexual orientation or gender identity,* **trade-union membership** *and activities*, **and the processing of genetic** *or biometric* **data or data concerning health or sex life***, administrative sanctions, judgments,* **criminal** *or suspected offences,* **convictions or related security measures shall be prohibited.** **2. Paragraph 1 shall not apply** *if one of the following applies*: **…** **(h) processing of data concerning health is necessary for health purposes and subject to the conditions and safeguards referred to in Article 81; or** **(i) processing is necessary for historical, statistical or scientific research purposes subject to the conditions and safeguards referred to in Article 83; or** *(ia) processing is necessary for archive services subject to the conditions and safeguards referred to in Article 83a;*

Purpose limitation	**Article 6:** **1. Member States shall provide that personal data must be:** **..** **(b) collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes. Further processing of data for historical, statistical or scientific purposes shall not be considered as incompatible provided that Member States provide appropriate safeguards;** **..** **(e) kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the data were collected or for which they are further processed. Member States shall lay down appropriate safeguards for personal data stored for longer periods for historical, statistical or scientific use.**	***Article 6*** **1. Processing of personal data shall be lawful only if and to the extent that at least one of the following applies:** **(a) the data subject has given consent to the processing of their personal data for one or more specific purposes;** **(b) processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract;** **…** **2. Processing of personal data which is necessary for the purposes of historical, statistical or scientific research shall be lawful subject to the conditions and safeguards referred to in Article 83.** **…** *4. Where the purpose of further processing is not compatible with the one for which the personal data have been collected, the processing must have a legal basis at least in one of the grounds referred to in points (a) to (e) of paragraph 1. This shall in particular apply to any change of terms and general conditions of a contract.*	***Article 6*** **1. Processing of personal data shall be lawful only if and to the extent that at least one of the following applies:** **(a) the data subject has given consent to the processing of their personal data for one or more specific purposes;** **(b) processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract;** **…** **2. Processing of personal data which is necessary for the purposes of historical, statistical or scientific research shall be lawful subject to the conditions and safeguards referred to in Article 83.**
	--	***Recital 40*** ***The processing of personal data for other purposes should be only allowed where the***	***Deleted***

		processing is compatible with those purposes for which the data have been initially collected, in particular where the processing is necessary for historical, statistical or scientific research purposes. Where the other purpose is not compatible with the initial one for which the data are collected, the controller should obtain the consent of the data subject for this other purpose or should base the processing on another legitimate ground for lawful processing, in particular where provided by Union law or the law of the Member State to which the controller is subject. In any case, the application of the principles set out by this Regulation and in particular the information of the data subject on those other purposes should be ensured.	
	Article 7 Member States shall provide that personal data may be processed only if .. (c) processing is necessary for compliance with a legal obligation to which the controller is subject;	Article 7(4): Consent shall ***not provide a legal basis for the processing, where there is a significant imbalance between the position of the data subject and the controller***.	Article 7(4): Consent shall ***be purpose-limited and shall lose its validity when the purpose ceases to exist or as soon as the processing of personal data is no longer necessary for carrying out the purpose for which they were originally collected. The execution of a contract or the provision of a service shall not be made conditional on the consent to the processing of data that is not necessary for the execution of the contract or the provision of the service pursuant to Article 6(1), point (b)***.
	Article 6 Member states shall provide that personal data must be: […]	Article 5 Personal data **must** be: …	Article 5 Personal data **shall** be: …

	(e) kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the data were collected or for which they are further processed. Member States shall lay down appropriate safeguards for personal data stored for longer periods for historical, statistical or scientific use.	(e) kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed; personal data may be stored for longer periods insofar as the data will be processed solely for historical, statistical or scientific research purposes in accordance with the rules and conditions of Article 83 and if a periodic review is carried out to assess the necessity to continue the storage;	(e) kept in a form which permits **direct or indirect** identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed; personal data may be stored for longer periods insofar as the data will be processed solely for historical, statistical or scientific research **or for archive** purposes in accordance with the rules and conditions of **Articles 81 and** 83 and if a periodic review is carried out to assess the necessity to continue the storage, **and if appropriate technical and organizational measures are put in place to limit access to the data only for these purposes** (**storage minimisation**);
Secondary processing and exemptions for research	Recital 34 Whereas Member States must also be authorized, when justified by grounds of important public interest, to derogate from the prohibition on processing sensitive categories of data where important reasons of public interest so justify in areas such as public health and social protection - especially in order to ensure the quality and cost-effectiveness of the procedures used for settling claims for benefits and services in the health insurance system - scientific research and government statistics; whereas it is incumbent on them, however, to provide specific and suitable safeguards so as to protect the fundamental rights and the privacy of individuals;	Recital 42 Derogating from the prohibition on processing sensitive categories of data should also be allowed if done by a law, and subject to suitable safeguards, so as to protect personal data and other fundamental rights, where grounds of public interest so justify and in particular for health purposes, including public health and social protection and the management of health-care services, especially in order to ensure the quality and cost-effectiveness of the procedures used for settling claims for benefits and services in the health insurance system, or for historical, statistical and scientific research purposes.	Recital 42 Derogating from the prohibition on processing sensitive categories of data should also be allowed if done by a law, and subject to suitable safeguards, so as to protect personal data and other fundamental rights, where grounds of public interest so justify and in particular for health purposes, including public health and social protection and the management of health-care services, especially in order to ensure the quality and cost-effectiveness of the procedures used for settling claims for benefits and services in the health insurance system, for historical, statistical and scientific research purposes, ***or for archive services*** .
	--	**Article 81(1b)** --	***Article 81(1b) - new*** ***1b. Where the data subject's consent is required for the processing of medical data***

			exclusively for public health purposes of scientific research, the consent may be given for one or more specific and similar researches. However, the data subject may withdraw the consent at any time.
	--	Article 81(2) Processing of personal data concerning health which is necessary for historical, statistical or scientific research purposes, ***such as patient registries set up for improving diagnoses and differentiating between similar types of diseases and preparing studies for therapies, is*** subject to the conditions and safeguards referred to in Article 83.	Article 81(2) Processing of personal data concerning health which is necessary for historical, statistical or scientific research purposes ***shall be permitted only with the consent of the data subject, and shall be*** subject to the conditions and safeguards referred to in Article 83. Article 81(2a) – **new** ***Member States law may provide for exceptions to the requirement of consent for research, as referred to in paragraph 2, with regard to research that serves a high public interest, if that research cannot possibly be carried out otherwise. The data in question shall be anonymised, or if that is not possible for the research purposes, pseudonymised under the highest technical standards, and all necessary measures shall be taken to prevent unwarranted re-identification of the data subjects. However, the data subject shall have the right to object at any time in accordance with Article 19***.
	--	--	Recital 123(a) – ***new*** ***The processing of personal data concerning health, as a special category of data, may be necessary for reasons of historical, statistical or scientific research. Therefore this Regulation***

			foresees an exemption from the requirement of consent in cases of research that serves a high public interest.
	Article 6: 1. Member States shall provide that personal data must be: (a) processed fairly and lawfully; (b) collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes. Further processing of data for historical, statistical or scientific purposes shall not be considered as incompatible provided that Member States provide appropriate safeguards;	Article 83: 1. ***Within the limits*** of this Regulation, personal data may be processed for historical, statistical or scientific research purposes only if: (a) these purposes cannot be otherwise fulfilled by processing data which does not permit or not any longer permit the identification of the data subject; (b) data enabling the attribution of information to an identified or identifiable data subject is kept separately from the other information ***as long as these purposes can be fulfilled in this manner***.	Article 83: 1. ***In accordance with the rules set out*** in this Regulation, personal data may be processed for historical, statistical or scientific research purposes only if: (a) these purposes cannot be otherwise fulfilled by processing data which does not permit or not any longer permit the identification of the data subject; (b) data enabling the attribution of information to an identified or identifiable data subject is kept separately from the other information ***under the highest technical standards, and all necessary measures are taken to prevent unwarranted re-identification of the data subjects***.