Legal Guidelines for TDM Practitioners

1. Introduction

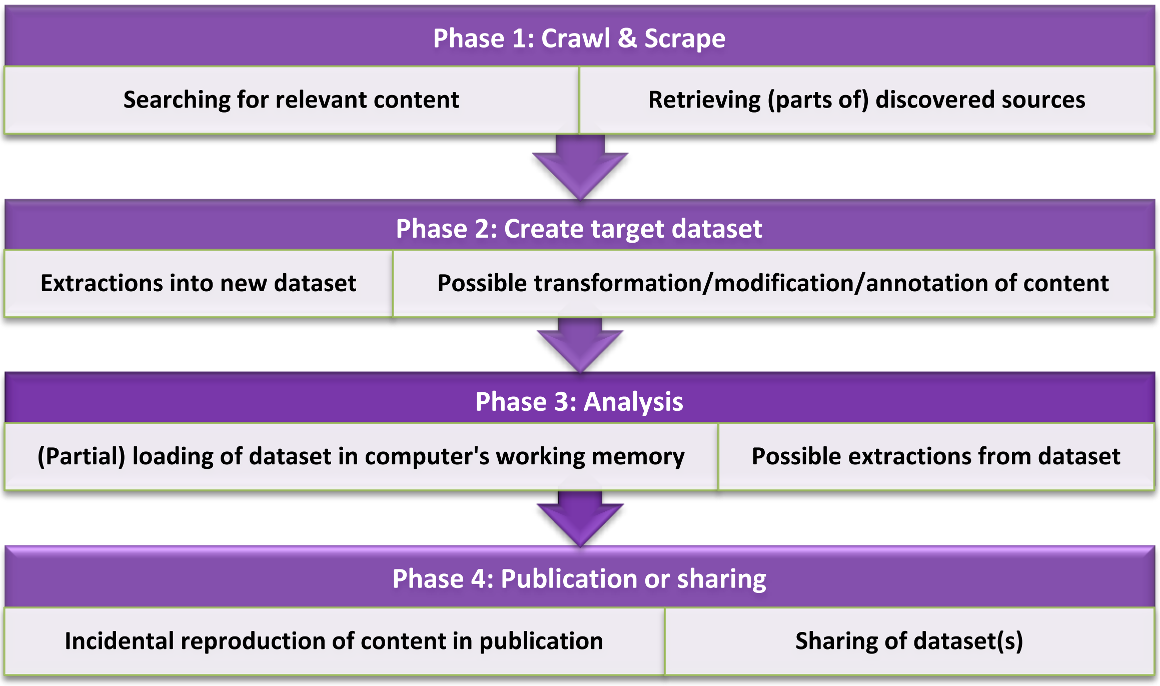
The legal landscape around TDM is more complicated than TDM practitioners may realise. In fact, there are many instances in which TDM is potentially unlawful. These guidelines are intended to give practitioners an overview of the legal landscape around TDM, so that they can be aware of potential legal issues and minimise legal risk.

In the absence of clear legal exceptions, intellectual property rights, including copyright, neighbouring rights, and *sui generis* database rights, will almost certainly be relevant when working with content created by others; these are discussed in section 3. TDM practitioners should also be careful to respect personal data and the privacy of any data subjects; data protection laws and best practices are discussed in section 4.

These guidelines are not intended to be comprehensive legal advice. Rather, they aim to give TDM practitioners a foundational overview of relevant legal considerations, to help understand when it might be necessary to seek expert legal advice.

2. Relevant legal considerations

To identify the legal risks of TDM, we first need to understand the activities involved in the TDM process. The outline in Figure 1 shows four general phases in the TDM process, with examples of acts that may be carried out in each phase. (Not everyone carrying out TDM will necessarily need to do all of these things, depending on the type of TDM and the purpose for which it is carried out.)
These phases will be referred to throughout these guidelines.



Phase 1: Crawl & Scrape

Searching for relevant content

Retrieving (parts of) discovered sources

Phase 2: Create target dataset

Extractions into new dataset

Possible transformation/modification/annotation of content

Phase 3: Analysis

(Partial) loading of dataset in computer's working memory

Possible extractions from dataset

Phase 4: Publication or sharing

Incidental reproduction of content in publication

Sharing of dataset(s)

Figure 1: Generalised overview of TDM processes

Before starting any TDM project, it is important to assess the potential legal issues of your project, and plan to avoid or minimise legal risks in your project design. For this purpose, some key questions you should ask about any TDM undertaking are:

- What sort of content am I going to use, and is it protected by or subject to any regulation?
- What sort of acts will I be carrying out on the content, and are these acts subject to specific rules under the relevant regulation?
- How should I deal with any applicable regulation to prevent or minimise the risk of my TDM project being rendered unlawful?
- In which cases should I turn to professional legal advice?

These guidelines should help you to carry out a rough evaluation of the legal risks of your TDM project, and to assess whether you should seek further legal advice.

When it comes to protected content, the two most common legal regimes that miners – at least in Europe – will face in practice are:

1. **Intellectual property rights**, more specifically copyrights, neighbouring rights and database rights
2. **Data protection rules**

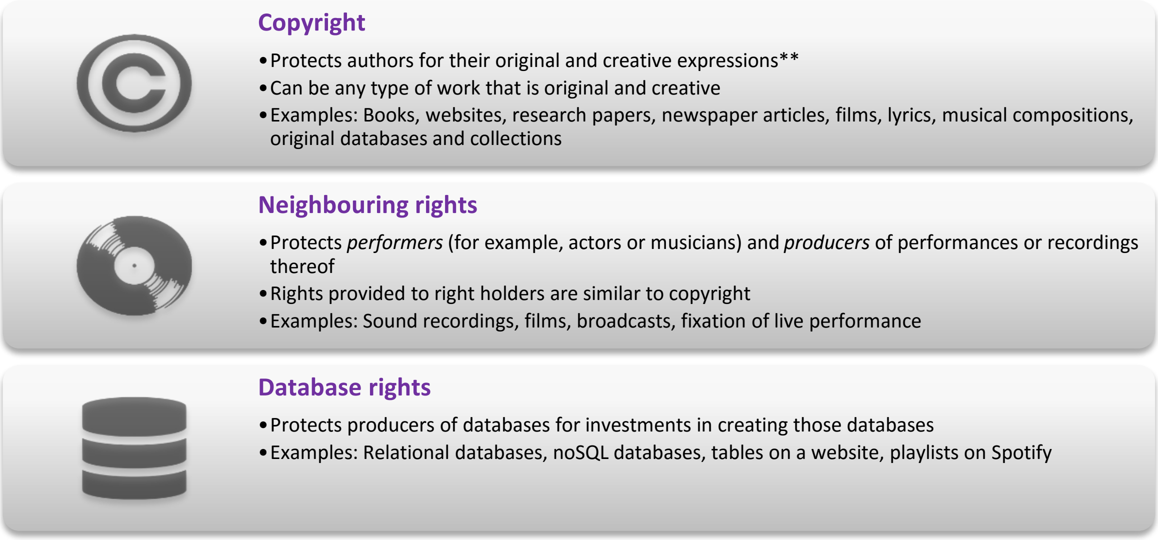
We will look at these regimes separately to help answer the questions posed above.

3. Mining others' intellectual property

Many TDM activities are carried out using content that is the intellectual property of other people, and subject to intellectual property (IP) rights.

When are IP rights relevant?

When mining content, there are three kinds of protection you need to consider: *copyrights, neighbouring rights* and *database rights*. These are the intellectual property rights that may be attached to the content you are intending to mine. It is important to establish whether any of these rights exist in the content you will be mining, because if they do, you might need permission from the right holders involved.



Copyright

- Protects authors for their original and creative expressions**
- Can be any type of work that is original and creative
- Examples: Books, websites, research papers, newspaper articles, films, lyrics, musical compositions, original databases and collections

Neighbouring rights

- Protects *performers* (for example, actors or musicians) and *producers* of performances or recordings thereof
- Rights provided to right holders are similar to copyright
- Examples: Sound recordings, films, broadcasts, fixation of live performance

Database rights

- Protects producers of databases for investments in creating those databases
- Examples: Relational databases, noSQL databases, tables on a website, playlists on Spotify

Figure 2: Intellectual property considerations for TDM

***Note that facts and data are not creative expressions, and do not attract copyright. Pure 'data mining' is therefore less likely to infringe copyright, except for the copyrights possibly existing in the collection of those data. Conversely, 'text mining' – including mining of other rich contents, such as images, films and music – is highly likely to be affected by copyright or neighbouring rights. In both text and data mining, you should always be aware of database rights in the collections of data, text or other contents.*

If you are dealing with any content similar to the examples in Figure 2, you should be aware that your TDM project could potentially infringe IP rights if you do not have permission from the rights holder. The following sections will help you evaluate whether you need to undertake further action.

Do I need to search for the rights holder?

If you have determined that you are dealing with protected content, you should establish what you are going to do with that content, and verify whether this is something that needs the consent of the IP rights holders. This is generally the case when you *copy*, either permanently or temporarily, or *publish* those contents in whole or in part.

- *Copying content:* In phases 1 to 3 of Figure 1, TDM activities usually involve making copies of content or (parts of) databases, ranging from retrieving copies from one or more sources, to transforming the contents into a (formalised) dataset that will be loaded into the computer's working memory when performing TDM analysis.
- *Publishing content:* If you are planning to share or disseminate any of your TDM results, or the underlying data or content sources, this is likely to be considered "publishing" - and will need permission in most instances as it relates to the exclusive rights of the rights holder to control the communication or redistribution of their content.

These acts need to be authorised by rights holders, unless special exceptions apply. Despite the existence of common European rules on copyrights and database rights, the applicability of and scope of these exceptions vary significantly across national borders. This means that if you work in multiple countries or collaborate with foreign colleagues, even within the EU, you will need to assess any relevant exceptions for each country you are operating in.[1]

As of April 2017, only the UK and France[2] have introduced exceptions in their laws that specifically allow you to use content for TDM without permission from rights holders. The UK exception only applies to copyright law, although a general non-commercial research exception exists for database rights in the UK. The French TDM exception applies to both copyright and database rights. In both countries, due to restrictions within the European Copyright Directive, these exceptions are limited to TDM for non-commercial and scientific research purposes where users have lawful access to content – for example because they have subscriptions to journals, or because they are freely available websites on the internet. These exceptions may benefit for example university researchers, whose research is for non-commercial scientific purposes. However, it is not entirely clear-cut when these *non-commercial* and *scientific research* conditions are met. For example, researchers involved in consortia with industry partners cannot be sure that they can benefit from such an exception.

In many European countries, other exceptions may also exist if you use content for:

- *Private and non-commercial purposes*: This may allow you to do text mining for your own private use.
- *Non-commercial research or teaching purposes in general*: some EU member states have an exception for certain acts carried out for research, some for teaching, and some for both. The scope of these is often very narrow and therefore unlikely to cover a full TDM process, if at all.
- *Temporary copies necessary to enable lawful use of a work*: This exception exists in all EU countries and may in many cases permit the part of the TDM process where the contents are temporarily loaded into the computer's working memory, although uncertainty exists regarding the extent to which this exception allows this.

[1] You can find an overview of implementations of exceptions at http://copyrightexceptions.eu.

[2] At the time of writing, a decree that would further detail the application of the French TDM exception was rejected by the Conseil D'Etat. Therefore the exception, despite being in the Intellectual Property Code, is not in force yet.

This project has received funding from the European Union's Horizon 2020 (H2020) Research and Innovation Programme.

These exceptions generally only permit TDM under either very specific circumstances, or one or a few phases in the TDM process.[3]

Step-by-step plan to minimise risk

To minimise risks, we advise you work through the following steps.

Step 1: Is it protected?
Establish whether the content to be mined is potentially protected by any copyrights, neighbouring rights or database rights. If yes, establish whether the corpus or whole body of contents is in the public domain, because all rights have lapsed.

Step 2: What am I going to do with it?
If you carry out any of the first three steps of the TDM process (Figure 1), you are likely to make copies that are subject to any right holders' approval. Such approval is also necessary when you publish or share TDM results, when these results contain original or modified versions of the contents you mined.

Public domain?

Copyright lasts 70 years after the death of the author. Historical sources may be out of copyright.
Neighbouring rights last 50 years after first publication, or 70 years in the case of phonograms.
Database rights last 15 years after publication. If a database is substantially modified, this term starts again counting from the day the modified version is published. Database rights can apply even to content that is out of copyright.

Retrieval from databases

If you retrieve information from a database, this will not infringe any database rights if you only retrieve an insubstantial part. Retrieving substantial parts – at once or bit by bit – of the database as a whole **does** affect the rights holders' exclusive rights.

Step 3: For what purpose?
Approval is not necessary when your activities are subject to an exception. For example, in some European countries, this may be the case when you make reproductions (such as those in steps 1 to 3) for non-commercial private or research purposes. No exception will apply if you share the full set of contents that you mined, but quoting from works in, for example, a research paper might be permitted in many European (and other) countries. Further, the sharing of facts and aggregated data (such as statistical representations), and new knowledge (such as newly created semantic annotations for TDM) always remains free if no original content is being shared.

Step 4: Do I have or need a licence?
In most cases, especially outside of a non-commercial private or research context, you cannot rely on exceptions to IP rights within Europe. Therefore, you should check whether you have an appropriate licence to mine the content. Even when you might benefit from an exception, that exception might be overridable by the terms of your contract. Therefore: always check your licences!

Please see our Guidelines for Licensees for more information.

Step 5: Do I need further legal assistance?
It is always better to be safe than sorry. If you have any doubt whether you:

- deal with protected sources,
- can rely on any exception, or
- should have a licence,

[3] For a detailed review of laws and policies affecting TDM, see FutureTDM Deliverable D3.3 (PDF)

This project has received funding from the European Union's Horizon 2020 (H2020) Research and Innovation Programme.

please consult an expert within your organisation, or seek advice from an external expert. If you are a TDM user who belongs to an academic or other institution, your library is likely to be the best starting point to understand what licensing conditions apply to content your institution has subscribed to. It might be even safer to take this step before going through the other steps!

Further materials

European harmonising directives

- Copyright & Neighbouring Rights: Copyright Directive 2001/29/EC: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32001L0029
- Database Rights: Database Directive 96/9/EC: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A31996L0009
    - And its national transitions: http://eur-lex.europa.eu/legal-content/EN/NIM/?uri=CELEX:31996L0009

Information on national copyright laws

- National copyright exceptions: http://copyrightexceptions.eu
- Sources for national IP legislation: http://www.wipo.int/wipolex/en/
- UK IPO on copyright and research: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/375954/Research.pdf
- Tool for calculating if works are out of copyright: http://outofcopyright.eu

Other

- FutureTDM deliverable elaborating on legal barriers: http://www.futuretdm.eu/knowledge-library/?b5-file=2374&b5-folder=2227

This project has received funding from the European Union's Horizon 2020 (H2020) Research and Innovation Programme.

4. Mining personal data

Important!

These guidelines alone are not sufficient to tell you how you should work with personal data, as this must be assessed carefully on a case-by-case basis. The guidelines are rather meant as an introduction to the principles and duties of data protection law. We always recommend you integrate data protection principles in the whole design of your TDM project, and always consult a data protection expert within or from outside of your organisation before you commence any TDM project involving personal data.

When is data protection relevant?

In Europe, you have to comply with specific regulations when you are dealing with (or 'processing') personal data. This means that when you mine any data relating to individuals, you should be aware of the rights and duties that come with it. Personal data is any data that relates to an identified or identifiable living[4] person, and can cover any sort of data as long as it enables you to directly or indirectly identify an individual. It also includes opinions about living individuals.

Examples of personal data

- Name, age, gender
- Home address
- Phone number
- Personal email
- IP address
- Bank account data
- Passport data
- Genetic data
- Health data
- Criminal records

You should also be aware that anonymised data can sometimes be de-anonymised by combining it with data from other sources. That is, if you hold an anonymised dataset, that data can become personal data again if new data is added to it which would allow you to identify individuals. Particularly in an online environment, where data from many different sources is increasingly combined, true anonymisation may be practically impossible.

Virtually *anything* you do with personal data is bound by European data protection rules, ranging from collecting and storing data to modifying or removing them. Therefore, if you deal with personal data in any of the phases in the TDM process (see Figure 1**Error! Reference source not found.**), you will need to comply with data protection law.

Principles and duties of European data protection law

Collection and further use

For the purpose of these guidelines, we distinguish three types of data use in the context of mining:

1. *Collection of personal data*: Retrieving any personal data directly from individuals or other sources (re-use of data).
2. *Use of personal data*: mining by you, someone else within your organisation, or on your behalf, of the retrieved data.
3. *Transfer of data*: transferring data to other parties.

Data minimisation vs. maximisation

There is a peculiar contradiction between the *data maximisation* (collecting and using as much data as possible) goal that makes big data and TDM so valuable, and the *data minimisation* principle of data protection regulation. The data minimisation principle entails the following:

- Personal data should only be collected for specified, explicit and legitimate purposes.
- Further use of data should be carried out in a manner compatible with the purposes for which they were collected (purpose limitation).

[4] Be aware that specific (self-)regulation or codes in some countries or sectors may also cover deceased persons.

This project has received funding from the European Union's Horizon 2020 (H2020) Research and Innovation Programme.

- Any use must be adequate, relevant and limited to what is necessary for those purposes.

Rights and duties

Personal data may only be processed on the basis of one of the following legal grounds:

- *Consent*: the person (data subject) to which the data relates has given their consent for the specified purposes.
- *Contract*: the use of the data is necessary to comply with a contract to which the data subject is party.
- *Legitimate interest*: you have a legitimate interest in using the data, which overrides the fundamental rights and interests of the data subjects, although public sector bodies may not rely on this anymore from May 2018.
- *Compliance with legal obligations*, *protection of the vital interests of the data subject*, or *performance of public interest task by official authority*: these grounds will generally not be relevant in the context of TDM.

Consent

Within the context of data protection law, consent by the data subject must be:

- Unambiguous: no doubt may exist
- Informed: all relevant information must be given to give informed consent
- Registered properly, in able to prove and review the consent from each individual afterwards

Other duties:

- Notify the relevant data protection authority that you process personal data. This *general* obligation will be abolished as of May 2018, and be replaced by procedures and mechanisms that rather focus on types of data use involving high risks. For example, notifications will have to be made in case of data breaches.
- Inform data subjects of your activities, if they are not already informed.

Rights of the data subject:

- Right to be informed
- Right to access their data
- Right to object to use of their data

Sensitive data

- racial or ethnic origin
- political opinions
- religious or philosophical beliefs
- or trade union membership,
- genetic data
- biometric data for the purpose of uniquely identifying a natural person
- data concerning health
- data concerning a natural person's sex life or sexual orientation

Mining sensitive data

European data protection law has a stricter regime for dealing with *sensitive* data. This is generally prohibited, unless you have legal grounds.

This project has received funding from the European Union's Horizon 2020 (H2020) Research and Innovation Programme.

Special provisions for research

On several aspects, the data protection framework provides for a lighter regime when personal data is used for *scientific or historical research* purposes. We give a few examples.

Purpose limitation and storage

Data must be processed for no other purposes than those for which the individual has given their consent to. With scientific research, however, it is often not possible to fully identify the purposes for which personal data are collected. Here the data protection framework has some leeway for scientific research: Further processing of collected data for scientific or historical research purposes will be considered to be compatible with the initial purposes for which the data is collected. Further, where data may normally be stored no longer than necessary for these initial purposes, longer storage is permitted when solely for scientific or historical research.

Data not collected from individuals

When data is re-used from other sources, the TDM researcher has not collected the data from the individuals themselves. Normally in such cases, they will have to inform all involved data subjects on the use of the data. However, in the particular case of using data for the purpose of scientific or historical research, a TDM researcher will not have to do this if it would be impossible or require disproportionate effort.

Right to be "forgotten"

Data subjects have a right to obtain the erasure of their personal data – to be "forgotten" – for example when the storing of their data is no longer necessary for the purposes for which it was collected or otherwise processed. This does not apply where the use of personal data is necessary for scientific or historical research purposes.

Safeguards

When personal data is being used for scientific or historical research, this research must be bound by appropriate safeguards, to respect the rights and freedoms of data subjects. If identification of data subjects is no longer necessary to fulfil the research purposes, this data shall be used in a manner that does not enable identification (through *pseudonymisation* or *anonymisation* of the data).

What is "research"?

It is not entirely clear what exactly constitutes "scientific" or "historical" research. However, when you carry out academic research adhering to academic standards, it will be more likely that this is regarded to be scientific research, than research carried out in an industrial context.

Be aware of special rules

When you use personal data in your scientific research, be sure to check whether in your research domain, or for the type of data you use, particular (domain-specific) regulation, self-regulation or codes of conduct apply. Such special rules commonly exist for the use of medical data or patient records.

Example: Health-data-driven research

Genetic data is considered *sensitive data*, and therefore carrying out TDM on genetic data can be problematic. One UK-based genetic research initiative[a] has taken the following steps to ensure they comply with data protection regulations:

- Data is anonymised: No data or names in files have a relationship with the names of patients
- Integrity and security of data is protected with encryption and physical locks on hard drives
- An in-house data protection officer oversees security, anonymisation and subsequent use of data

This example demonstrates good practice for those frequently dealing with data, for TDM or other purposes, with a dedicated expert to evaluate whether data is processed, stored and used legally and securely.

[a] This pseudonymised initiative within a UK health institution was the basis for a case study carried out within the BYTE project; see their *Case study reports on positive and negative externalities, p. 109*, available at http://new.byte-project.eu/wp-content/uploads/2015/06/FINAL_BYTE-D3-2-Case-studies-report-1-1.pdf

Do’s and don’ts

We cannot provide general guidelines on how each TDM project should deal with personal data, since this largely depends on the scale, nature and purpose of the TDM activities, as well as on the nature and source of the personal data. Dealing with data protection law and ethics is very complex and we therefore strongly recommend you always consult an expert in this area when designing your TDM project. This section provides lists of *do’s* and *don’ts* to give you some guidance as to the most important aspects of dealing with personal data in your TDM project.

Do’s

- Establish if you will use or mine personal data and whether it also includes sensitive data
- Assign a Data Protection Officer if TDM is one your organisation’s core activities, or if your organisation does TDM on a regular basis
- *Impact Assessment* (IA): establish what data you will use for what purposes, and who will have access to the data within and outside your organisation, and whether your use of personal data brings any high risks
- Check whether you have the legal grounds to collect and/or use the personal data
- *Privacy by design*: based on your IA, design your whole TDM project in a way that guarantees that you can safely and adequately use the personal data
- Look into sector-specific regulation, or self-regulation and codes of conduct within your domain, which may provide you more guidance and certainty on what you can do
- Anonymise data, so you are not dealing with personal data any more. Note that if you pseudonymise personal data, this is will still be personal data if the use of additional information enables you to attribute the data to a natural person
- Make every effort to ensure that the personal data you hold is accurate

Don’ts

- Only think of data protection issues when you actually start to mine
- Collect data and just assume that it does not concern any personal data
- Store and retain all data just because it may be useful in the future
- Randomly transfer or provide access to any data to third parties
- Re-use data from one project in another one, without making sure this is compatible with data protection rules, *even though* you had made sure that the use in the first project was compatible
- Share any personal data with the public, without proper consultation
- Make decisions affecting the data subject based solely on automated processing of their personal data – this is prohibited
- Ignore data subjects’ requests to access, rectify or erase data
- Transfer data outside the EU

Further materials

European legislation

- Data Protection Directive 95/46/EC: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:31995L0046
  - This directive does not apply directly to European citizens and organisations, but is implemented in the national laws of its member states
- General Data Protection Regulation: https://publications.europa.eu/en/publication-detail/-/publication/3e485e15-11bd-11e6-ba9a-01aa75ed71a1/language-en
  - Will apply from May 25, 2018
  - From that date, will repeal the above-mentioned Data Protection Directive

Institutions

- List of national Data Protection Authorities: http://ec.europa.eu/justice/data-protection/article-29/structure/data-protection-authorities/index_en.htm
- Glossary of data protection terms: https://edps.europa.eu/data-protection/data-protection/glossary_en
- UK ICO *Guide to data protection*: https://ico.org.uk/for-organisations/guide-to-data-protection/
- UK ICO *Big data, artificial intelligence, machine learning and data protection*: https://ico.org.uk/media/for-organisations/documents/1541/big-data-and-data-protection.pdf

Other

- FutureTDM deliverable elaborating on legal barriers: http://www.futuretdm.eu/knowledge-library/?b5-file=2374&b5-folder=2227

This project has received funding from the European Union's Horizon 2020 (H2020) Research and Innovation Programme.