GeoERA
INFORMATION
PLATFORM

Deliverable

A study of the risks associated with geodata delivery in Europe

Authors and affiliation:

Christopher Luton, BGS

E-mail of lead author:
chto@bgs.ac.uk

Version: 1.0

Deliverable Data		
Deliverable number	D10.4	
Dissemination level	Public	
Deliverable name	A study of the risks associated with geodata delivery in Europe	
Work package	WP10, IPR and data policy issues	
Lead WP/Deliverable beneficiary	BGS	
Deliverable status		
Submitted (Author(s))	8/10/2021	Christopher Luton
Approved (Coordinator)	28/10/2021	Jørgen Tulstrup

GENERAL INTRODUCTION

This document looks at right to use and publish data with a particular emphasis on new initiatives and legislation and how it affects the processes. From initial observation of data and software by scientists at an early stage of the research process, it looks at the key considerations to enable either open access to output data or the required licenses to make the dissemination legal.

EXECUTIVE REPORT SUMMARY

The European Union is very keen to see publicly supported research outputs in Europe and the rest of the world put to social and economic benefit. If not, the data can often be simply archived and remain hidden. Unfortunately, copyright law seems to act as a real disincentive to publish and make data openly available, although if the processes are correctly followed, consent overrides the need to worry about potential breaches. But this then requires the active scientist/researcher to find the originator of "the copyright works", which is often not easy, especially bearing in mind the long periods within which copyright can subsist. To try to facilitate better access to materials, the EU has introduced FAIR which it hopes will accelerate the process through what it calls a "FAIRification" process.

Running alongside these constrains on access to data is the worry about breaching personal data rights, the so-called GDPR. Always in the back of the minds of scientists needs to be the requirement to consider whether GDPR will be breached, as fines/penalties for mere copyright breaches is low, where breaches of GDPR and personal data can be very serious and costly – both financially and reputational-wise.

When permission has been sought for the use of copyright material, the next consideration is the best type of license that is used. Creative Commons has generally been accepted (CC:BY) as the means of disseminating most results for the GeoERA platform.

TABLE OF CONTENTS

1. Background

At the commencement of the GeoERA programme it was assumed that a number of the GeoERA science groups would be reliant on using the GeoERA Information platform to ensure that what they were producing or adapting would be available for open/public use. The reasoning was that it would be somewhat pointless for datasets and other data products to be developed by the GeoERA scientific teams, with no likelihood they could be ever be used: there was a need to go right back to the beginning of the product development process to ensure that the necessary consents/freedoms were in place and that the scientists understood not only the legal hurdles, but what they needed to put in place to ensure free access to the data and data products.

For this reason, a questionnaire was developed (attached at Annex A) as an aide memoir for the scientists, so when they gained access to the data "building blocks", they knew what they needed to take into consideration and put in place before embarking on scientific development so, wherever possible, free access and dissemination could be made of the resulting outputs.

It was a little understood that in most countries trade and business secrets were an important consideration, however what was in question here was legal constraints on use – so if the trade secret has no legal protection, should it be free to use?

2. Some Basic Assumptions about Data

WP10 has discussed a lot in earlier WP deliverables about what data is and what it actually means. The Questionnaire at Annex A also delves into data and data types, because it is only after understanding what data holdings are involved that the legal position can be fully assessed as to whether data is free to use without permission and/or what the restriction are in reality. An example of this is the concept of "derived data", which is effectively trying to determine if a new copyright is originated when existing "raw" data becomes so "derived" or removed from the original that it cannot be considered the source data. It becomes in effect a completely new originated set of data with new original ownership.

3. Raw data

The starting point is raw data. By raw data we mean the original data that has been collected from a source and not yet processed or analysed. Raw data provides the foundation for any downstream analysis. In many cases the captured or collected data may be unique and impossible to reproduce, such as time points in weather measurements and interviews. For this

reason, they should be safeguarded from any possible loss. Moreover, raw data will typically be lossless - i.e. those file formats that are not compressed such as TIFF files for image data as opposed to compressed JPEG file format.

Finally, in some cases, raw data may have additional information that may be specific to a brand and/or type of instrument used to capture the data. For example, Leica microscopes use a proprietary data format but is also a container for lossless data - the container contains metadata specific to the Leica microscopes that allows reading, writing and analysis through Leica software.

So, what do we mean by processed and analysed data?

By processed data we mean data that has already undergone some kind of intervention. For instance, the data have been digitised, compressed, translated, transcribed, cleaned, validated, checked and/or anonymised.

By analysed data we mean data already processed, interpreted and analysed. Analysed data can assume several representations (text, tables, graphs, etc.), in order to facilitate a better understanding and communication of the data.

In most cases, one can also consider raw data as the official data, that is, the master copy of any given record (sometimes called a golden copy). As well as providing the starting point for derivatives generated downstream through analyses, there may be additional branches from which this data is used for other analyses. Therefore, in a typical workflow, the recommendation is that a scientist creates a copy of the raw data which one can use as a "working copy". The original data should then be archived in an appropriate manner for long-term preservation. The working copy can then be used for processing and analysing without worrying about overwriting.

However, it must also be understood that not all research data are digital. Most researchers keep handwritten laboratory notebooks, journals and other materials, examples of which may be surveys, paintings, fossils, minerals and tissue. However, non-digital data can be converted to a digital source in a variety of ways.

The general point to make as far as derived data is concerned and what a scientist can safely use depends on a number of factors. Clearly the original data, the processed and analysed data all

move the newly developed data slight further away from what can be called the original copyright material. The Questionnaire at Annex explains the route for a scientist to take and at what point it is safe to assume that a newly "derived" copyright has been formed.

4. Copyright Law

Without a doubt the main obstacle to disseminating or releasing data/data product, etc is copyright. Copyright owners of data and data products have a right in law to protect their innovative developments. Copyright law is largely harmonized in the European Union, although country to country differences exist. Copyright in the (EU) Union is furthermore dependent on international conventions to which the European Union is a member, such as the TRIPS Agreement and conventions to which all Member States are parties (such as the Berne Convention). The body of law was implemented in the EU through a number of Directives, which the member states needed to enact into their national law. The main copyright directives are the Copyright Term Directive, the Information Society Directive and the Directive on Copyright in the Digital Single Market.

The Directive on Copyright in the Digital Single Market, formally the Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC is a European Union (EU) directive which has been adopted and came into force on 7 June 2019. It is intended to ensure "a well-functioning marketplace for copyright". It extends existing European Union copyright law and is a component of the EU's Digital Single Market project. The European Council (EC) describes their key goals with the Directive as protecting press publications; reducing the "value gap" between the profits made by Internet platforms and by content creators; encouraging collaboration between these two groups, and creating a copyright exception for text and data mining.

The directive was introduced by the European Parliament Committee on Legal Affairs on 20 June 2018, and a revised proposal was approved by the parliament on 12 September 2018. The final version, which resulted from negotiations during formal trilogue meetings was presented to the parliament on 13 February 2019. The measure was approved by on 26 March 2019 and the directive was approved by the Council of the European Union on 15 April 2019. Member states had two years to pass appropriate legislation to meet the Directive's requirements, so it is now seen as good law throughout Europe.

The directive has generally been opposed by major tech companies and most Internet users, as well as human rights advocates, but supported by media groups and conglomerates, including newspapers and publishers. Two of the Directive's articles have drawn significant discussion. Draft Article 11 (Article 15 of the directive), known as the "link tax", gives newspapers more direct control and re-use of their work, which may impact some Internet services like news aggregators. Draft Article 13 (Article 17 of the directive) tasks service providers that host user-generated content to employ "effective and proportionate" measures to prevent users from violating copyright. Tech companies expressed concern that this would necessitate the need for upload filters.

A broad concern with the Directive is on the use of fair dealing through the directive, and that it could quell freedom of speech. "Fair dealing" is a limitation and exception to the exclusive right granted by copyright law to the author of a creative work. Fair dealing is found in many of the common law jurisdictions of the Commonwealth of Nations, and each country appears to have its own slight variations on how the law should be applied.

Fair dealing is an enumerated set of possible defences against an action for infringement of an exclusive right of copyright. Unlike the related United States doctrine of "fair use", fair dealing cannot apply to any act which does not fall within one of these categories, although common law courts in some jurisdictions are less stringent than others in this regard. In practice, however, such courts might rule that actions with a commercial character, which might be naïvely assumed to fall into one of these categories, were in fact infringements of copyright, as fair dealing is not as flexible a concept as the American concept of fair use.

In the United Kingdom, as an example, under the Copyright, Designs and Patents Act 1988 (CDPA), fair dealing is limited to the following purposes:

1. research and private study (note: both must be non-commercial);
2. criticism / review / quotation, and news reporting; and
3. as well as parody, caricature and pastiche and illustration for teaching.

Although not actually defined as a fair dealing, incidental inclusion of a copyrighted work in an artistic work, sound recording, film, broadcast or cable programme does not infringe copyright. Since 2014 the UK has protected the fair dealing exceptions from override by contracts or contractual terms and conditions.

Contrary to the often-stated view, the provisions of the CDPA do *not* state the amount of an in-copyright work that may be copied for the purposes of non-commercial research or private study or to single copies of the work, where the copies are made by the researcher or student himself. Such restrictions only apply to copies made by or on behalf of a librarian, or by a person, other than the researcher or student himself, who knows or has reason to believe that "it will result in copies of substantially the same material being provided to more than one person at substantially the same time and for substantially the same purpose". This is reflected in other European States and is the sort of grey area within which lawyers work. Hence the text in the Questionnaire that advises that there are no hard and fast rules to copyright breaches and it is all a case of negotiation and quantum and quality of use.

For copying beyond the boundaries of fair dealing, universities and schools in the UK obtain licences from a national copyright collective, the UK Copyright Licensing Agency (CLA). Under these licences, multiple copies of portions of copyrighted works can be made for educational purposes. In fact, the UK was the only Member State of the European Union that did not have a private copy exception.

5. Permissions for use of GeoERA Scientific Information

Generally copyright law clearly imposes severe restrictions on what use can be made of third party data/information. One of the best, and legally sufficient, ways of overriding this restriction is by seeking express permission, which is a fundamental requirement of the Questionnaire developed. How that permission is sought and what that permission entails will be different in each case: some owners of copyright might be selective in where any of their materials are used or published, some may have concerns about adaption of their materials (perhaps into something not intended), other may insist on acknowledgement. Of course, copyright does not subsist "legally" in all materials, however universally throughout the world, a relatively clear definition of what is copyright and what it subsists in, has been established. For any user of materials within which copyright might subsist, it is probably better that permission is sought rather than to get into lengthy legal wars over whether copyright is present in any material.

6. Derived Materials

Deriving materials, which was discussed earlier in this text, with a new copyright clearly has its merits, however to create a completely derived product or process that is sufficiently separated

or different to the original to be considered legally original in its own right is not so easy. The test for this is difficult to determine and sometimes even when one might have been merely guided by the original, it is still considered original and in the ownership of the originator.

From a legal standpoint, a scientific material user needs to consider what are termed "non-copy derived" and "copy derived" data or products.

Non-copy derived means that the derived data set does not include a copy of the whole or any substantial part of the original information (as defined by copyright legislation and case law) and the derived data set cannot be reverse engineered to create a copy of the original information, or any substantial part of it.

Copy derived means that the derived data set includes a copy of the whole or any substantial part of the original information or that the derived data set can be reverse engineered to create a copy of the original information or any substantial part of it.

If it is "copy derived" you will generally be breaching copyright law.

As pointed out above, if an external body owns the copyright in the materials/information and one is engaged in copying that information, you will be seen as *prima facie* breaching copyright. However, as a general rule what the law states is:

- Firstly, there must be sufficient objective similarity between the infringing work and the copyright work, or a substantial part thereof, for the former to be properly described, not necessarily as identical with, but as a reproduction or adaption of the latter; and
- Secondly, the copyright work must be the source from which the infringing work is derived.

So, a key consideration is substantiality/quality. Not every act of copying is infringement: one must have copied the whole work or a substantial part of it. If we think about a "substantial part" we need to think of quality and not just quantity, as a single figure or image if sufficiently important (whatever its size) as a part of the copyright work could be considered substantive!

An example of this might be a single image in a large book produced for commercial use. If the single image explains the whole text of the book in a diagram, this would be considered a breach under the substantive/quality concept.

7. FAIR Guiding Principles

The European Commission is very keen to see research data published and wants to see the use of copyright law limited, particular where Government funded public/university research outputs can lead to a better standard of social and human welfare. The FAIR principles describe how research outputs should be organised so they can be more easily accessed, understood, exchanged and reused. Major funding bodies, including the European Commission, therefore promote FAIR data to maximise the integrity and impact of their research investment. The EC supports FAIR data not as a standard but as a framework to follow when designing a Data Management Plan. As such, it has produced a set of guidelines for FAIR data management.

It is intended that when scientists are carrying out project work within GeoERA they consider the principles of FAIR – which is now recognised by the EU as an efficient method of data management/stewardship, allowing for enhanced knowledge discovery and innovation. There is in fact a growing "Go FAIR Community" being established which hold regular meetings.

Good data stewardship is the key to knowledge discovery and innovation. To generate value for a research community beyond the initial researchers, funding agencies are increasingly setting requirements for proper data stewardship of research data. Beyond proper collection, annotation, and archival, data stewardship includes the 'long-term care' of research data, with the goal that they can be found and re-used in downstream studies and research. To facilitate good data stewardship, a broad community of international stakeholders have hence developed the FAIR Data principles, which have been embraced by both the European Commission and the G20. The first formal publication of the FAIR Principles further describes the rationale behind them.

Since its beginning in early 2018, the "GO FAIR" community has been working towards implementations of the FAIR Guiding Principles. This collective effort has resulted in a three-point framework that formulates the essential steps towards the end goal, a global Internet of FAIR Data and Services where data are Findable, Accessible, Interoperable and Reusable (FAIR) for machines. In short, the FAIR Data Principles propose that all scholarly output should be:

- **Findable:** easy to identify and find for both humans and computers, with metadata that facilitate searching for specific datasets,

- **Accessible:** stored for long term so that they can easily be accessed and/or downloaded with well-defined access conditions, whether at the level of metadata, or at the level of the actual data,

- **Interoperable:** ready to be combined with other datasets by humans or computers, without ambiguities in the meanings of terms and values,

- **Reusable:** ready to be used for future research and to be further processed using computational methods. This requires adequate information about how the data were obtained and processed (provenance) and an appropriate license.

Set up around April 2020 there are working groups developing methods, tools and documentation around the Three-point Framework:

- Metadata 4 Machines working group

- FAIR Implementation Profile working group

- FAIR Data Point working group

Driven by the urgency of the COVID-19 pandemic, the 3-point FAIRification Framework is under active development in multiple projects. The immediate objective of the three working groups is the creation of a handbook consolidating the methods and resources for running M4M workshops, for building FAIR Implementation Profiles and for setting up FAIR Data Points.

The FAIR Data Principles apply to metadata, data, and supporting infrastructure (e.g., search engines). Most of the requirements for findability and accessibility can be achieved at the metadata level. Interoperability and reuse require more efforts at the data level. The scheme below depicts the FAIRification process adopted by GO FAIR, focusing on data, but also indicating the required work for metadata:
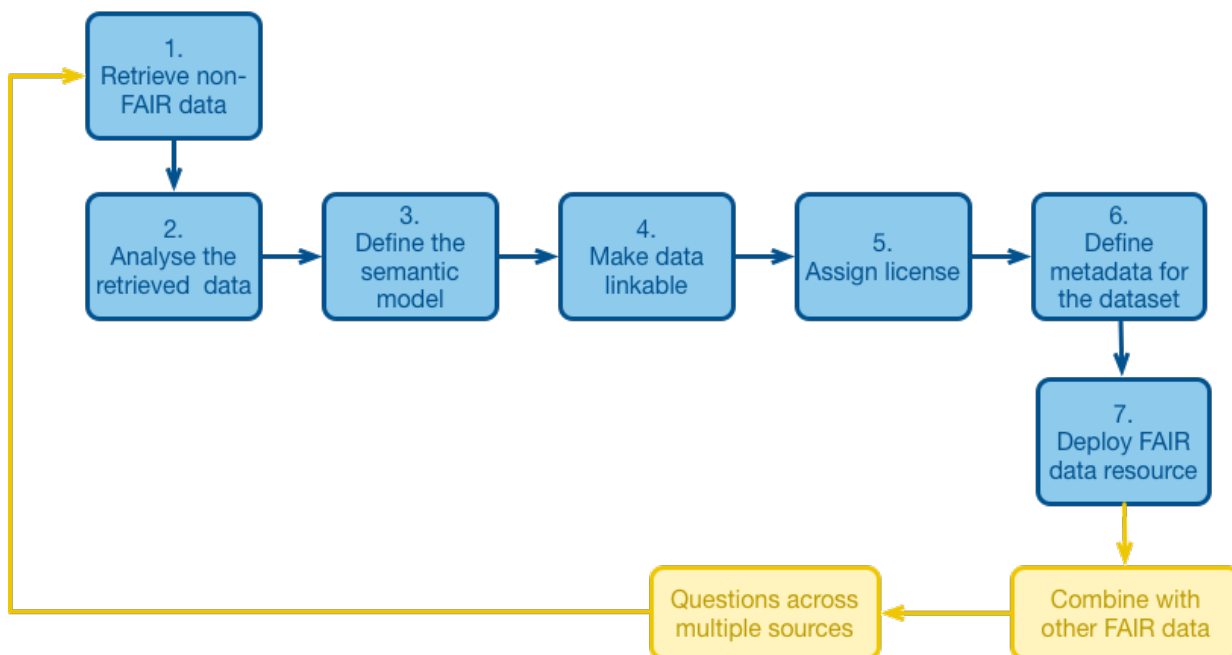
The FAIRification process consists of the following steps:

1. **Retrieve non-FAIR data:** gain access to the data to be FAIRified.

2. **Analyse the retrieved data:** inspect the content of the data: Which concepts are represented? What is the structure of the data? What are the relations between the data elements? Different data distributions require different methods for identification and analysis. For instance, if the dataset is in a relational database, the relational schema provides information about the dataset structure, the types involved (the field names), cardinality, etc.

3. **Define the semantic model:** define a 'semantic model' for the dataset, which describes the meaning of entities and relations in the dataset accurately, unambiguously, and in a computer-actionable way. Depending on the dataset, defining a proper semantic model may require a significant effort, even for experienced data modellers. A good semantic model should represent a consensus view in a particular domain, for a particular purpose. Therefore, it is good practice to search for existing models. Semantic models often contain multiple terms from existing ontologies and vocabularies. A vocabulary is a computer-readable file that captures terms, their URIs, and descriptions. An ontology can be roughly described as a vocabulary with hierarchies, meaningful relations among concepts, and their constraints. These

conceptual models allow us to classify our data models and data items using the provided terms, concepts, and conceptual structures.

4. **Make data linkable:** The non-FAIR data can be transformed into linkable data by applying the semantic model defined in step 3. Currently, this is done using Semantic Web and Linked Data technologies. This step promotes interoperability and reuse, facilitating the integration of the data with other types of data and systems. However, the user should evaluate the feasibility of this step for the given data. It is a sensible thing to do for many types of data (e.g., structured data), but it may not be relevant for other types (e.g., the pixels or audio elements in images, audio data, and videos). Of course, the annotations about the images, audio, and video (e.g., data about identified regions of images, or about parts of an audio file) could very well be made linkable.

5. **Assign license:** Although license information is part of the metadata, it is important to incorporate the licence assignment as a separate step in the FAIRification process to highlight its importance. The absence of an explicit license may prevent others to reuse data, even if the data is intended to be open access. Clarity of licensing status will become more important with automated searches involving more licensing considerations. The conditions under which the data can be used should be clear to machines and humans.

6. **Define metadata for the dataset:** As explained by many of the FAIR Principles, proper and rich metadata support all aspects of FAIR.

7. **Deploy FAIR data resource:** deploy or publish the FAIRified data, together with relevant metadata and a license, so that the metadata can be indexed by search engines and the data can be accessed, even if authentication and authorisation are required.

Of interest note there is no "O" for "Open" in FAIR. Proponents of FAIR data often also stress that data should be as open as possible, access only being restricted where necessary. Note here that FAIR data and open data are different, although there are similarities. The key difference is that open data should be available to everyone to access, use, and share, without licences, copyright, or patents. It is expected that open data at most should be subject to attribution/share-alike licenses.

FAIR data, however, uses the term "Accessible" to mean accessible by appropriate people, at an appropriate time, in an appropriate way. This means that data can be FAIR when it is private, when it is accessible by a defined group of people, or when it is accessible by everyone (open data). It depends completely on the purpose of the data, where the data currently is in its lifecycle, and the end-usage of the data. For example, new experimental data may only be accessible by the generator and their group to start, then with consortia partners as the findings become refined, and finally with the public upon publication.

Personally sensitive data may never be publicly accessible and usable. Commercially sensitive data may be held privately for stretches of time after collection and interpretation. Users are also free to use more restrictive licenses to govern how the data may be reused.

8. GDPR/Personal Data

GDPR is adequately covered off in the Questionnaire. It runs alongside copyright in being a requirement to check and for one which a scientist needs to make a provision. Personal data breaches are serious and regulators in EU States are becoming far more aware of breaches. Since the introduction of the General Data Protection Regulation (EU) 2016/679 (GDPR), and probably in earlier national legislation, there is clearly a lot more concern in Europe over the controlling, processing, etc of personal data. So when accessing or using Materials special consideration need to be given to the presence of or inclusion of personal data. When it is present it needs to be carefully managed and processed according to law.

Although the GDPR creates heightened obligations for entities that process personal data, it also creates new exemptions for research as part of its mandate to facilitate a Digital Single Market across the EU. Specifically, the GDPR exempts research from the principles of storage limitation and purpose limitation so as to allow researchers to further process personal data beyond the purposes for which they were first collected. Research may furnish a legitimate basis for processing without a data subject's consent.

The Regulation also allows researchers to process sensitive data and, in limited circumstances, to transfer personal data to third countries that do not provide an adequate level of protection. To benefit from these exemptions, researchers must implement appropriate safeguards, in keeping with recognized ethical standards, that lower the risks of research for the rights of individuals.

9. Conclusion

WP10 has looked a range of issues concerning the legal right to use and publish research outputs. The most serious of these is copyright (the right of the originator/owner to stop use and publication), and the WP10 deliverables have looked at exceptions to copyright and also ways in which derived data can be used to avoid breaches.

The start for any scientists has to be the use of the Questionnaire. This looks at trying to focus the mind of the scientist on the key areas to think about. How easy is "consent" available to be sought from a copyright owner? Does the scientist know who the originator is of a set piece of work? Is it worth taking the risk the owner will never be found? All these types of questions need to be going through the mind of a scientist wanting to use third part materials: a scientist should always err on the side of caution, because copying another person's work can have its consequences. Not necessarily in terms of the fine or financial penalty (which may only be small), but more in terms of the reputation of the scientist and his/her institution.

Breaching GDPR/personal data rights is a more serious matter and the penalties/fines imposed can be very serious. Repeat offences can lead to the fines being multiplied. In most cases a breach will occur when personal data is processed and, for example, a data originator's details and contained within a set of published data and the penalty will fall back on various parties, including perhaps the publishing scientist's organisation. Of course, many breaches may be able to be settled with an apology, but even with an apology it may not stop a regulator issuing a serious fine.

It is argued within the scientific and legal profession that the GDPR will not hinder "Open Science" in the long run. Those who have handled data correctly in the past will not have to learn a lot of new things. Ultimately, the GDPR will make sure that research can and will be done openly and fairly because its transparency rules make backroom research unviable. The reasons are:

(i)   the GDPR is only relevant for scientific data from human participants (i.e., irrelevant for most scientific disciplines); and

(ii)   the GDPR is only relevant for personal data. If the data provided by human subjects is anonymous (e.g., most of psychology's data), the GDPR does not apply.

Most institutions have counteracted the GDPR issues by:

- revamping consent forms. In case of personal data which cannot be anonymized, we ask for consent to share it;
- making better checks whether data sets really are anonymous before making them open; and
- by trying to improve operational security (e.g., safer handling of pseudonym lists).

In the scientific field in, for example, Germany there is a feeling that GDPR does not worsens the data availability situation. They believe that personal data can be made available if the value to society is greater than privacy concerns for subjects. As a reverse scenario, they believe publishing open data about people has become "easier".

This deliverable has discussed FAIR in some detail, however the important point is that the FAIRification process established and promotes the use of licences, which really goes back to the main point about *copyright* being the key consideration for any scientist wanting to use and publish third party data.

The last written set of legal advice for WP10 looked at types of licences which should be adopted for use by GeoERA scientists. The GeoERA teams decided that for the safe dissemination of research results the scientists should opt for Creative Commons licenses which are widely used throughout the world. The licencing processes have been developed to take into consideration all the copyright legal issues and their website even allows one to choose what type of licence they will use at https://chooser-beta.creativecommons.org/

GeoERA has decided, after much debate, to generally use the CC:BY licence. This license allows re-users to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator. The license allows for commercial use and is accompanied by the following logo.

For licensing software, GeoERA has adopted a choice of licences **The Creative Commons Public Domain Dedication** which is used by US Government bodies

CC0 (aka CC Zero) is a public dedication tool, which allows creators to give up their copyright and put their works into the worldwide public domain. CC0 allows re-users to distribute, remix, adapt, and build upon the material in any medium or format, with no conditions.

The **GNU General Public License** (**GNU GPL** or simply **GPL**) is also available and is a series of widely used free software licences that guarantee end users the freedom to run, study, share, and modify the software. The licenses (the GPL series) are all copyleft licenses, which means that any derivative work must be distributed under the same or equivalent license terms. This is in distinction to permissive software licenses. GPL was the first copyleft license for general use. Historically, the GPL license family has been one of the most popular software licenses in the free and open-source software.

Annex 1 Questionnaire

<u>**Introduction**</u>

This questionnaire has been designed as an aid to support enquiries as regards freedom to use third party data, data products, and software ("Materials")*. Note in some countries trade and business secrets are an important consideration, however I believe what we are talking about here is legal constraints on use – so if the trade secret has no legal protection it should be free to use.

The Questionnaire does not provide an absolute or definitive answer, but guides you as to some the questions felt relevant when accessing and using materials that originate from third parties. It is suggested you use one form for each dataset to avoid confusion.

The format of the questions allows for a tested process probably best described by answering four questions:-

- Are the Materials free from any intellectual property considerations? (copyright, etc)
- Are there any terms and conditions attached or associated with use of the Materials?
- To what extent can the Materials be used without breaching legal rights?
- If I develop a new product or process, under what type of licence can I release it? Can it be an "Open" product/process?

The second area you will need to think about (Schedule 1) is whether whatever you develop or create is sufficiently derived or different to the Materials that you have been using. If it is, you might create a completely derived product or process that is sufficiently separated or different to the original to be considered legally original in its own right. However the test for this is not easy to determine and sometimes even when you have been merely guided by the original, it is still considered original and in the ownership of the originator.

It is intended that when scientists are carrying out project work within GeoERA they consider the principles of FAIR – which is now recognised by the EU as an efficient method of data management/stewardship, allowing for enhanced knowledge discovery and innovation. More details are at Schedule 2.

The last consideration is that of personal data. Since the introduction of the General Data Protection Regulation (EU) 2016/679 (GDPR), and probably in earlier national legislation, there is a lot more concern in Europe over the controlling, processing, etc. of personal data. So when accessing or using Materials special consideration need to be given to the presence of or inclusion of personal data. When it is present it needs to be carefully managed and processed according to law. More guidance is provided at Schedule 3.

		Question	Comment	Answer
1		Have you identified materials not owned by your organisation that you wish to use in connection with your work? (Note: if the materials are from your own organisation you may also need permission).	Third party information and data from non-Government Sector Organisations (GSOs) that is intended to be used for GeoERA projects should be identified. GSOs may have restrictions on the use of their own data and information (e.g. only after publication has taken place, embargo periods, etc.). Such limitations need to be added (as at point 9 below)	(this response can be a "yes" or "no" or might result in a list of materials).
2		What do the materials comprise?	A short description is required. Please use an additional page if further explanation is required	
3		Where are the materials available from and how did you find them – known repository, Government link, university, open website?	Were possible please provide a reference number and identify any DOI or acknowledgement required	
3		From the location of the materials, was there any mention of who owned the materials?	We probably mean here the copyright ownership – however some companies/organisations licence in materials so they don't actually own them. Bear in mind that if you have obtained materials from an organisation that has itself licensed them in from elsewhere, it may be that the owner of	

		the materials, rather than the organisation you have obtained them from, will need to be consulted.	
5	What is your intended use of the materials and will the finished product/dataset include a substantial part of the materials? Could the materials be reverse engineered from your product?	Copyright law allows one to derive materials as long as you don't copy a substantial part of the original. If the source materials are not visible in what you create and if you cannot reverse engineer from what you have created to the original material, this is termed *non-copy derived* material. Further information regarding *copy derived* and *non-copy derived* is contained at Schedule 1.	
6	Do you intend to modify the format of the materials in any way?	I.e. are you planning to convert the materials to a different delivery format; and/or are you converting from a digital to a 'flat' format (or vice versa)	
7	Do you see a need to share the materials, or anything that you derive from the materials, with anyone else in connection with this project?		
8	Do you see any future commercial applications for what you have derived from the materials?		

9	Were the materials accompanied in any way with terms and conditions concerning use or are they freely available (without any cost) to everybody using FAIR Principles (Schedule 2) both during and post the GeoERA project?	This could take many forms: Creative Commons licences, Open Access, conditions concerning commercial and non-commercial use, free use but improvements made open, etc. If terms and conditions apply please provide links or describe these in detail for all restricted Materials.	
10	Are you sure that these terms and conditions of use cover all of your proposed usage of the materials (i.e. questions 5-8 above)?	If you are not sure, or if you are sure that some (or all) of your proposed usage of the materials is outside of what is permitted under the terms of use you have seen, you need to contact the owner of the materials to get full permission. You also need to check not only that you have permission, but whether or not the owner of the materials may have a claim of IPR ownership over anything you derive from their materials (this is more likely if you are producing *copy derived* materials – see Schedule 1). This may have implications for whatever you have derived: i.e. could the owner request a royalty/revenue share in relation to any commercial use that you maybe intending to make?	

11	Processing personal data/information. Does any of the materials you wish to use contain "Personal Data"? Examples might be where personal names have been left in the digital content. Please see Schedule 2 that covers the obligations that might be on you to ensure this matter is clearly accounted for and that where you pass on personal data instructions are given as to what and cannot be processed by the user.	Obligations on you will relate to whether you are "controlling" or "processing" personal data.	
12	When passing on data that holds Personal information you will need to consider further you and your users' position. Are you passing on personal data?	If you transfer or pass on the personal data, you may have a requirement to ensure that anyone you pass the personal data to also complies with the Data Protection Laws/Regulations. See Schedule 3.	

Schedule 1

Non-copy derived means that the derived data set does not include a copy of the whole or any substantial part of the original information (as defined by Copyright legislation and case law) and the derived data set cannot be reverse engineered to create a copy of the original information, or any substantial part of it.

Copy derived means that the derived data set includes a copy of the whole or any substantial part of the original information or that the derived data set can be reverse engineered to create a copy of the original information or any substantial part of it.

If an external body owns the copyright in the materials/information and you are engaged in copying that information, you are *prima facie* breaching copyright. However as a general rule what the law states is that:-

- Firstly, there must be sufficient objective similarity between the infringing work and the copyright work, or a substantial part thereof, for the former to be properly described , not necessarily as identical with, but as a reproduction or adaption of the latter; and
- Secondly, the copyright work must be the source from which the infringing work is derived.

So, a key consideration is substantiality/quality. Not every act of copying is infringement: one must have copied the whole work or a substantial part of it. When we think about a "substantial part" we need to think of quality and not just quantity, as a single figure or image if sufficiently important (whatever its size) as a part of the copyright work could be considered substantive!

Schedule 2

FAIR Data: Background and Rationale

Good data stewardship is the key to knowledge discovery and innovation. To generate value for a research community beyond the initial researchers, funding agencies are increasingly setting requirements for proper data stewardship of research data. Beyond proper collection, annotation, and archival, data stewardship includes the 'long-term care' of research data, with the goal that they can be found and re-used in downstream studies and research. To facilitate good data stewardship, a broad community of international stakeholders have developed the FAIR Data principles, which have been embraced by both the European Commission and the G20. The first formal publication of the FAIR Principles further describes the rationale behind them.

FAIR Data

The paper 'The FAIR Guiding Principles for scientific data management and stewardship' was written in 2016 and is the first formal publication of the FAIR principles. In short, the FAIR Data Principles propose that all scholarly output should be:

- **Findable:** easy to identify and find for both humans and computers, with metadata that facilitate searching for specific datasets,
- **Accessible:** stored for long term so that they can easily be accessed and/or downloaded with well-defined access conditions, whether at the level of metadata, or at the level of the actual data,
- **Interoperable:** ready to be combined with other datasets by humans or computers, without ambiguities in the meanings of terms and values,
- **Reusable:** ready to be used for future research and to be further processed using computational methods. This requires adequate information about how the data were obtained and processed (provenance) and an appropriate license

Please consult 'The FAIR Data principles explained' for a more detailed description.

Of interest note there is no "O" for "Open" in FAIR. Proponents of FAIR data often also stress that data should be as open as possible, access only being restricted where necessary.

Further details are at:-

https://www.go-fair.org/fair-principles/

Schedule 3

As background, Data Protection Laws means the General Data Protection Regulation (EU) 2016/679 (GDPR) and any other law applicable to Europe relating to the protection of personal data and the privacy of individuals, including where applicable guidance and codes of practice issued by any European Authority.

Understanding whether you are processing personal data is critical to understanding whether the law applies to your activities. "*Personal Data*" means personal data as defined in the legislation, which is information that relates to a *living* identified or identifiable individual. What identifies an individual could be as simple as a name or a number or could include other identifiers such as an IP address or a cookie identifier, or other factors.

If it is possible to identify an individual directly from the information you are processing, then that information may be *personal data*. If you cannot directly identify an individual from that information, then you need to consider whether the individual is still identifiable. You should take into account the information you are processing together with all the means reasonably likely to be used by either you or any other person to identify that individual.
Even if an individual is identified or identifiable, directly or indirectly, from the data you are processing, it is **not** personal data unless it 'relates to' the individual.

When considering whether information 'relates to' an individual, you need to take into account a range of factors, including the content of the information, the purpose or purposes for which you are processing it and the likely impact or effect of that processing on the individual. It is possible that the same information is personal data for one controller's purposes but is not personal data for the purposes of another controller.

Information which has had identifiers removed or replaced in order to pseudonymise the data is still personal data for the purposes of the law (GDPR).
Information which is truly anonymous is not covered by the law (GDPR). If information that seems to relate to a particular individual is inaccurate (i.e. it is factually incorrect or is about a different individual), the information is still personal data, as it relates to that individual.

For the purposes of this Questionnaire, "**controller**", "**data subject**" and "**personal data**" have the meanings given under the Regulation; "**processing**" has the meaning given under the Regulation (and "**process**", "**processed**" and "**processes**" shall be construed accordingly); and "**Regulation**" means the General Data Protection Regulation (EU) 2016/679, as defined as part of the Data Protection Laws.

To the extent that you provide another ("user") with *personal data* in connection with the performance of any agreement or arrangement with the user, the user must agree that, unless the parties specifically agree otherwise, each (you and the user) shall be independent controllers of the *personal data* in their own right. The user must agree they shall comply with all applicable data protection and privacy laws in relation to the *personal data*, including the Regulation. If required, you may wish to ensure a potential user provides details of any *personal data* it intends to process in accordance with the questions below:-

PROCESSING OF PERSONAL DATA

This series of instructions for your user sets out the scope, nature and purpose of the processing of Personal Data by the user on your behalf and constitutes your written instructions.

1. THE SCOPE, NATURE AND PURPOSE OF PROCESSING

    1.1 User may process Personal Data in order to provide a service to its users.

2. THE DURATION OF PROCESSING

    2.1 User may process Personal Data as permitted by you for a duration (to be determined).

3. THE TYPES OF PERSONAL DATA

    3.1 User may process Personal Data of the following types:
        (a)   [personal details]
        (b)   [financial details]
        (c)   [employment and education details]
        (d)   [goods or services provided]

    3.2 User may not process sensitive Personal Data of the following types:

        (a)   [physical or mental health details]
        (b)   [racial or ethnic origin]
        (c)   [religious or philosophical beliefs]
        (d)   [trade union membership]
        (e)   [political opinions]
        (f)   [genetic data]
        (g)   [biometric data]
        (h)   [sex life and sexual orientation]]

4. THE CATEGORIES OF DATA SUBJECT

    4.1 User may process Personal Data relating your

        (a) [employees]
        (b) customers and clients]
        (c) [suppliers and service providers]
        (d) [advisors, consultants and other professional experts]
        (e) [complainants and enquirers].