Open Data, Law and Licensing

Open Data, Law and Licensing[1]

Francis Davey[2]

Introduction

Law

Some previous attendees of this course have given feedback that the subtle interplay between court decisions and the law was not something they were familiar with. So this section offers some very brief notes on the point. Readers who are already familiar with the law in general (including lawyers) should jump to the next section.

Jurisdiction

The legal world is, broadly speaking, divided into "jurisdictions". In each jurisdiction the law is different and interpreted by different sets of courts. In the UK the jurisdictions are:

- England and Wales
- Scotland
- Northern Ireland

Scots law is much less like the law of the rest of the UK, but it still has a great deal in common.

Some legislation in the UK is UK-wide: for example copyright and immigration law. Other legislation is specific to only parts of the UK (eg Scotland and England and Wales share the same employment law).

[1] version 4.4

[2] contact@francisdavey.co.uk

Some countries have one jurisdiction for the whole country – eg France. In other countries quite the reverse is true. As well as the UK, a good example is the US where there are 51 different jurisdictions (one per state plus the District of Columbia).

In the UK there is also an overlapping jurisdiction, that of the European Union. Where EU law and UK law conflict, EU law takes precedence.

Sources of Law

Law consists mostly of a combination of two things.

First written statements of the law by legislators. In the UK these include Acts of Parliament (also known as "statutes") as well as a motley collection of what is known as "secondary" legislation with titles such as regulations, rules, orders and so on. Most (but not all) forms of secondary legislation are made in a standard form known as a "statutory instrument" which requires unique numbering and publication.

In Scotland the Scottish Parliament can also make its own "Acts of the Scottish Parliament".

The other significant component of law is the interpretation of the law by the courts.[3] In the UK a great deal of law was once made entirely by the courts working from various first principles and without any legislation to go on. For example the law of breach of confidence was entirely made up by the courts of England and Wales without help from any legislators. On the other hand, copyright and database rights are entirely based in legislation, though they have been subsequently interpreted by the courts in some cases in great detail.

A substantial amount of the law relating to databases is part of European law. For our purposes, European law (the law of the European Union) consists primarily of three forms of legislation: the treaties setting up the EU, regulations and directives. In theory regulations are laws that apply directly in member states and directives are laws that need to be "transposed" into the law of a member state. The idea being that a directive will set out a template for the law, but some detail will be left to the legislators in each member state to fill in.

For example, the Database Directive[4] was transposed into UK law by the Copyright and Rights in Databases Regulations 1997.[5]

[3] In Scots law there are some significant academic works by the so-called "institutional writers" which are treated as being formal statements of law.

[4] Directive 96/9/EC

[5] See: http://www.legislation.gov.uk/uksi/1997/3032/contents/made

Precedent

In the UK, courts adopt a system known as "precedent". This arranges the courts into a hierarchy. If a court higher up the hierarchy makes a decision about a point of law, courts lower in the hierarchy are bound by that decision and cannot depart from it. Courts at the same level will usually follow earlier decisions of courts at their level (though the law on whether they have to do so is extremely complex and not easy to describe in a short series of notes).

In England and Wales the hierarchy is:

1. The Supreme Court (formerly the House of Lords)
2. The Court of Appeal
3. The High Court (and some superior tribunals)

Below the high court are various other courts (magistrates' courts and county courts) which are not sufficiently important to set precedents at all.

Courts will still take seriously decisions made by other courts which they respect.

The European Court of Justice

The European Court of Justice (ECJ or sometimes now CJEU for Court of Justice of the European Union) has an unusual role. It can adjudicate disputes between member states and individuals with the EU, but its most significant role is in answering questions referred to it by courts in the Member States.

For example, in the *Fixtures Marketing* case (see below), a court in Athens asked the ECJ three questions about the database right. The ECJ answered them and left the final decision on the case to the court in Athens.

The ECJ is often asked to decide questions that depend on the particular facts of the case and it routinely makes clear that it is up to the national court to make those decisions. All the ECJ does is to clarify points of EU law that are referred to it.

The ECJ's decisions on EU law bind all courts in the Member States. It does not consider itself bound by its own previous decisions but it would be unusual for it to depart from a clear line of case law it had established.

Ownership of data

Strictly speaking it is impossible to "own" data in the sense of having a property right in factual information. But there are now two property rights that apply to collections of data, technically known as "databases".

Copyright

Some history Copyright began in Great Britain as a protection for "books and other writings". It restricted the right to copy the books or other writings (hence the term "copyright") to the authors of the books for 14 years from publication.

Over time copyright protection was extended to many other forms of expression, for example paintings, sculptures, plays and photographs. These are collectively known as "works". As protection extended it became unclear what exactly copyright protected. It was generally understood that copyright shouldn't be able to protect facts. But that it should be able to protect adaptation of a work, eg translations into another language, dramatizations of a literary work. But it was not clear precisely where the line should be drawn

For example, there were some early cases where judges were unsure about whether maps could be protected since they simply represent what is already out there. There were also questions over drawings of plants and (later) photographs. Eventually it was accepted that copyright could protect all these kinds of work.

The courts converged on a distinction between "ideas" – which were not protected – and the "expression" of those ideas – which was protected. This is known as the "idea/expression" dichotomy. You could take facts recorded in someone's work, provided you did not copy the expression of the work.

Sometimes this caused difficulties where the work was, say, a table of information or a compilation. The courts settled on the idea that a compilation of facts might be protected by copyright even if the facts themselves were not.

In English law the concept of "originality" acted as a kind of gate-keeper to prevent copyright being too extensive. If a work was not original, there was no copyright in it.

At first it wasn't clear what kind of originality was required. In particular was it enough to put a lot of effort into a work where that work was not, necessarily, particularly creative. For example, creating a directory of streets in London requires a lot of effort, but it is not particularly original since anyone could go out and do it. Should copyright protect that data collection effort?

European harmonisation

All that argument, though very interesting for academics, is now past. In Europe, whether or not a work is protected by copyright (the "subsistence" of copyright) has been harmonized by three directives:

- the software directive – for computer programs

- the database directive – for databases
- the information society directive – for everything else.

In all cases the criterion (replacing the English conception of "originality") is that a work is protected if it is its author's "own intellectual creation".[6] The ECJ has explained that "own intellectual creation" implies a creative choice made by the author. They must make use of "formative freedom" so that the author puts the stamp of their personality on work. If there is no "formative freedom" – eg if there is only really one way to express something – then there can be no copyright.

Even under the new EU criterion a lot of things are capable of being protected by copyright. For example some (sufficiently creative) newspaper headlines might be protected.

How copyright protects the structure of databases will be dealt with next, but open data may involve copyright works as the **contents** of a database. In an open data context there is probably a contrast between:

- pure data – eg temperature values or GPS co-ordinates (almost certainly not protected individually)
- copyrightable works – eg images, user comments

General features of copyright

Some useful general features of copyright are worth noting:

- no formalities are required – copyright "just happens" when a work is created, you do not have to register or claim it in any way[7]
- you do not even have to mark your work with a copyright notice (C) or something similar
- copyright belongs to the author in most cases (films and sound recordings being an exception)
- **unless** the author creates the work in the course of employment, when it will first belong to the employer

[6] Some English courts seem to have been slow to catch up with this and you will still find decisions talking about "originality". There are some subtleties here – particularly for photographs and images – but for our purposes we can treat the only criterion that applies as "own intellectual creation".

[7] In the US, registration of copyright is optional, if a work that has been registered is infringed, the owner may be able to obtain a much higher award of damages for the infringement.

Infringement of copyright

Copyright in a work is infringed by doing any of the following without permission of the copyright owner:

- reproducing the work (i.e. copying it)
- distributing the work
- communicating the work to the public (eg publishing it on a web page or via an API)

Modern database protection

For the purpose of the database directive, a "database" is

"a collection of independent works, data or other materials which (a) are arranged in a systematic or methodical way, and (b) are individually accessible by electronic or other means"

Thus a "database" is what a computer scientist might call a "data set". A database does not have to be held on a computer. Examples of databases which are not electronic might be:

- a library (arranged according to the Dewey Decimal system)
- an anthology of poems organised by author (and date)
- a parish register (arranged by date)

If a database is held on a computer, the computer software that manages the database is not a part of the "database" for IP purposes. It would usually be protected by copyright law as a computer program. Copyright operates slightly differently for computer programs than it does for databases and is probably off-topic for this handout.

Database right

I deal with database right first, because it was the first of the two rights in databases to be dealt with by the European Court of Justice.

Database right is quite unlike copyright. It aims not to reward authorship, but to reward *investment*. It was created by the European Union in a, possibly mistaken, hope that it would encourage investment in the creation of a database infrastructure. The database right is sometimes referred to in EU circles as the "sui generis right" because it is not like any other kind of right.

The database right is also unknown in the US and generally internationally. This means that a database protected by database right in the EU will not enjoy that protection in the US. This has important implications for European institutions publishing databases to the internet. It means that a US based institution might be able to appropriate the database and use it in the US without worrying about database right.

The database right applies to a database where there has been a substantial investment in one or more of three factors:

- obtaining
- verifying
- presenting

the contents of the database. Here "substantial" can mean either quantitatively substantial – for example lots of time or money; or qualitatively substantial – for example the use of experts to collect the data (as happened in *Football Dataco v Stan James below*).

Fixtures Marketing v Organismos prognostikon agonon podosfairou[8] The Court of Justice explained that "obtaining" was different from "creating". The football leagues of England and Scotland claimed they had a database right in the list of football "fixtures", i.e. which teams would play each other on particular dates and in particular venues. The court decided that what the leagues were doing was "creating" data. It was they that decided whether or not two teams would play one another. They could not be said to be "obtaining" it.

This conclusion surprised quite a few commentators. Databases like the football league's fixtures list may even have been in lawmakers' minds when the database directive was drafted. The court of justice explained that the purpose of the database right was to encourage the investment in creating database infrastructure. The football league needed no encouragement via the database right to invest in their database of fixtures lists. They would have created such a thing anyway.

Football Dataco v Stan James[9] Where is the line drawn between "obtaining" and "creating"? Some academics have suggested that a common form of data collection – going out and measuring data (eg meteorological data, mapping data for Open Street Map) – was really "creating" the data on the theory

[8] C-444/02 (European Court of Justice)

[9] [2013] EWCA Civ 27, followed in *Calor v Flogas*[2013] EWHC 3060 (Ch) where even the parts of Calor's customer database that were created anew by Calor were thought to be protected

that the data doesn't exist before it is actually recorded. The Court of Appeal of England and Wales has rejected this view in no uncertain terms. It is not certain that the Court of Justice would take the same view, but the question hasn't been asked them. For the moment a substantial investment in measuring data would give the database right in English law.

Apis-Hristovich v Lakorda C-545/07 This was a case about databases of public legal information. For an open data perspective, the most interesting part of the decision was that simply because information is collected form publicly accessible data sources does not preclude it from being protected by the database right.[10]

The court (re)-iterated a few other points that are worth remembering:

- "Extraction" (see below) relies on a transfer of the contents of the database. The purpose of that transfer is irrelevant.
- It may be that the contents of database B resemble those of an earlier database A. That may be because of a transfer of the contents of A to B but there may be other good explanations. There is no rule B infringes A just because it looks the same.

Infringement

The database right protects the owner against three things being done without their permission. The first two are:

- extraction; or
- re-utilisation

of a substantial part of the database.

But the authors of the directive were also clearly worried about the possibility of taking "little and often". The directive also makes conduct an infringement of the database right if it consists of all of the following:

- repeated and systematic;
- extraction and/or re-utilisation of **insubstantial** parts of the contents of the database;
- implying acts which:

[10] Indeed publicly created data might even be subject to the database right of the public body that is curating it, see http://www.techdirt.com/articles/20130211/08050521945/europes-database-right-could-throttle-open-data-moves-there.shtml

- conflict with a normal exploitation of that database
- or which unreasonably prejudice the legitimate interests of the maker[11]

An **extraction** is involved when the database appears in a new medium. It does not matter if there has been human intervention – even selective intervention – in between.

For example in the Directmedia Case,[12] a University owned a database of the "100 most important poems in German literature between 1730 and 1900. Directmedia created their own anthology, using the database as a guide to which poems to select from that period. They were selective and did not use all the poems. The Court of Justice decided that could be an "extraction".

Re-utilisation means making (a substantial part of) the database available for members of the public to access.

For example in the Innoweb Case,[13] a meta search engine ("GasPedaal") would, in response to a user query, make its own queries to a number of automotive websites and in particular one known as Autotrack. A member of the public could, in principle, access a substantial part of Auotrack's database by using GasPedaal and hence the meta search engine's activities constituted "re-utilisation".

Derived works

What the database right does not seem to protect is a non-database work that has been created using a database – for example a graphic (such as a map or info-graphic) or a report summarising the results in a database. This is a subtle point because, if the derived work contains sufficient information that was in the original database to count as a "substantial part", then creating the derived work will be an "extraction", which would require permission of the database right owner.

But in many cases the derived work will not contain a copy of a substantial part of the information in the database and so there will be no extraction or re-utilization of the database.

This means that the degree to which a database right owner can use the database right to control the *use* to which their data is subsequently put is limited.

Copyright in databases

Copyright in a database is a right independent of any database right. In other words it is possible for either copyright or database right to exist in a database as well as both or neither. Neither right is "superior" to the other.

[11] Article 7(5)
[12] Directmedia Publishing v Albert-Ludwigs-Universität Freiburg C-304/07
[13] Innoweb v Wegener ICT Media BV C-202/12

To obtain copyright a database must be original, in that one of two things must be the author's "own intellectual creation". Either the:

- selection or
- arrangement.

For example the 100 most important poems in German literature between 1730 and 1900 could claim copyright on the basis of the selection of the poems to include. This would apply to many anthologies. Wherever what is and is not in the database is the result of the exercise of the author's "own intellectual creation".

Playlists and CD's

It has been widely reported that the Ministry of Sound have sued the music streaming company Spotify for infringement of copyright in their playlists. Oddly, the database directive specifically mentions this sort of situation in recital 19 which says:

"Whereas, as a rule, the compilation of several recordings of musical performances on a CD does not come within the scope of this Directive, both because, as a compilation, it does not meet the conditions for copyright protection and because it does not represent a substantial enough investment to be eligible under the sui generis right"

In other words: no database right or database copyright for you. The phrase "as a rule" means that there may be some exceptions, but the drafters of the directive clearly though that a typical CD could not normally attract either database right or database copyright.

Other rights

There are a few other rights that might protect information alone, which I will deal with briefly here.

- Confidential information – if information is given in confidence then it is a *breach of confidence* to communicate it to someone else. Information that has been published will almost never be confidential. The law of confidential information is fairly common-sense. For open data the take-home is to make sure that data you publish hasn't been obtained in breach of confidence or been the subject of confidence at some point.

- "Hot news"[14] – this is an idea that had some currency in the US for a while, though it is probably now much less important[15] – it was a doctrine that protected "hot" (i.e. time critical) news.
- German Leistungsschutzrecht für Presseverlege (aka "Lex Google") a right over news content for one year.

Effective technological measures

Database may also be protected by various forms of digital rights management, passwords or other forms of technology intended to control their use. These are known as "effective technical measures" in EU law. In many situations, circumventing an effective technological measure can be treated as the equivalent of an infringement of copyright and in some cases is a criminal offence.

It should be unusual to meet ETM's in the context of open data.

Open licensing

What is a licence?

Note that in UK English "licence" is a noun and "license" a verb. In US English both are "license". The US usage is very common online and some licences thus have names that look odd to an English lawyer's eye, eg the Creative Commons License.

A licence is just a permission to someone ("the licensee") to do something for which they would need the permission of someone else ("the licensor"). It need not – and often is not – be a contract.

For example: the postal worker who delivers our post walks up our driveway. If they did not have our permission to do so they would be a trespasser. The law implies a licence for lawful visitors to go to the front door of a house or building. This is clearly not a contract and could be revoked by telling a visitor to go away. The licence would not be implied where there was a clear notice forbidding entrance (eg "no hawkers").

In the context of copyright, a copyright licence gives permission to do one or more of the acts restricted by copyright (eg reproduction or communication to the public). A licence may have conditions attached. Breach of the conditions would withdraw the permission given by the licence. In many cases this would result in the licensee in breach of conditions of the licence infringing copyright.

Licences may also be restricted in various ways, eg they may be

[14] International News Service v. Associated Press, 248 U.S. 215 (1918)

[15] EFF: The "Hot News" Doctrine After Fly On the Wall: Surviving, But On Life Support (at https://www.eff.org/deeplinks/2011/06/hot-news-doctrine-surviving-life-support)

- exclusive or non-exclusive
- limited by jurisdiction (eg a UK licence)
- limited by time
- revocable in certain circumstances (or simply at the choice of the licensor)
- subject to the making of regular payments or payments calculated by some use made by the licensee (eg royalties)

Hence a common bit of boilerplate one sees is:

"a non-exclusive irrevocable, perpetual, royalty-free worldwide[16] licence …"

What is "open"?

The word "open" is used (or mis-used) by different groups to mean different things. For the ODI and for the purposes of today's talk, "open" follows the Open Definition:

A piece of data or content is open if anyone is free to use, reuse, and redistribute it — subject only, at most, to the requirement to attribute and/or share-alike."[17]

Share-alike / viral licensing

One of the pioneers of open licensing was Richard Stallman, the founder of the GNU project.[18] For Stallman it was important that users of software had access to the source code so they could make their own modifications and alterations to it. Those changes might then be used to create better software for everyone, not just the person who made the modifications. The problem was if he simply released the source code into the public domain, then anyone making modifications could keep any changes to themselves and not contribute back to the general community.

His solution was to create a licence that permitted others to use his software, along with the source code, but on condition that if they passed the software on to others – whether they sold it or not was irrelevant – they had to do make the source code available and do so under the same licence. The licence would then spread "virally" to all adaptations of the original.

Stallman used the term "free software" for his product. Not that it was free to buy, but that it was free for others to use a re-use. "Free as in speech, not as in beer" is how he put it.

[16] We are beginning to see (rightly in my view) the adoption of a practice long existing in Hollywood which is to spell out that the licence extends beyond planet Earth.

[17] See http://opendefinition.org/ The Open Definition is curated by the Open Knowledge Foundation (OKF). I declare an interest in that I frequently act for the OKF.

[18] GNU is a recursive acronym for Gnu's Not Unix. This is a computer science joke.

Examples:

- GPL (GNU Public Licence) – open source software
- CC-BY-SA (Creative Commons Attribution Share-Alike) – for works of art, literature etc
- ODbL (Open Database License) – for collections of data

There are many non-viral licences. For example:

- CC0
- CC-BY
- PPDL
- Open Government Licence

Multiple licences

Where a work is subject to more than one licence any user of that work will have to consider both licences and try to comply with any conditions they impose.

For share-alike restrictions, the rule is usually that the same or a "compatible" licence is used on any derived work. If a number of works – or in the case of data, data sources – are mixed together then there is a danger of **licence incompatibility**. This could make it difficult or impossible to mix together incompatibly licensed material.

A useful tool (which I haven't checked for legal accuracy) for discovering whether it is possible to mix licences can be found here at http://clipol.org/tools/compatibility. The ODI also makes available information about licence compatibility at https://github.com/theodi/open-data-licensing/blob/master/guides/licence-compatibility.md

For that reason, writing your own open licence from scratch is a bad idea. It is unlikely to be compatible with other open data licences and will almost certainly create more problems than it solves.

Rule 1 of open licensing
Never write your own open licence.

Which licence?

Remember that items of data may be protected by copyright. The whole collection of data may be protected by database copyright and database right. You

may want to use a different licence for the contents of your database than for the database as a whole. You might even have to do that if the contents were licensed to you under specific open licences.

Government – OGL (now in version 2.0[19])

Pure public domain

Both CC0 and the PDDL are strong statements that attempt to give up all rights in the material licenced. Both cover the database right as well as database copyright.

Creative Commons

Creative commons offers a number of options

- use restrictions: none, or non-commercial (NC)
- re-use restrictions: none, share-alike (SA) or no derivatives (ND)

All options (other than CC0 above) require attribution (BY). Which results in 6 possible variants:

- BY
- BY-ND
- BY-SA
- BY-NC
- BY-NC-ND
- BY-NC-SA

Both NC and ND are **not** open licence choices since they impose further restrictions on use. NC is particularly unhelpful because it is so unclear what counts as "commercial" use. Eg, does it apply to a website which is monetized by advertising? Thus only CC-BY and CC-BY-SA are "open" licences.

Creative Commons have recently introduced version 4.0 of their licence suite,[20] which now covers the database right (and analogous rights in other jurisdictions) as well as copyright. A CC licence may be applied to any licensed work – so it may be applied simply to the structure of a database and not its contents (in much the same way as ODC licences) – but unless otherwise noted a CC licence applied to a database is likely to apply to both the contents and the (structure, collection etc of the) database.

[19] See https://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/
[20] http://creativecommons.org/

Open Data Commons[21]

As well as the PDDL the ODC offers two database licences:

- ODC-By (attribution)
- ODC-ODbL (attribution and share-alike)

These apply only to the database and not its contents. If the contents may contain copyrightable material, you should consider an open licence for the contents as well – eg CC-BY or CC-BY-SA.

This means that both ODC-ODbL and CC-BY-SA v4.0 may be used for the share-alike licensing of a database.

Creative Commons v Open Data Commons?

CC and ODC's share-alike licences (CC-BY-SA and ODC-ODbL) do not operate in precisely the same way.

In principle, both try to impose a duty to share-alike (i.e. share under a suitable licence) any work derived from the licensed work, so that licensee are forced to share back into the commons the fruits of using the licensed work. But the two licences do so differently.

ODC-ODbL, contemplates three different kinds of derived work: "derivative databases", "collective databases" and "produced works":

- a "derivative database" is just a new database based on the licensed database (eg by modifying, translating or otherwise adapting it)
- a "collective database" is a larger database containing the licensed database in unmodified form but combined with other independent databases
- a "produced work" is some other work (eg a report or info-graphic) derived from the licensed database (or a derivative database or collective database containing the licensed database).

Publishing a "produced work" requires identification of the original licensed database (as part of attribution). If the produced work was produced from a derivative database requires that a copy of the derivative database (or a practical machine-readable copy of the differences) is published.

By contrast, CC relies on a concept it calls "adapted material". Adapted material is restricted to material that would have required the licensor's permission

[21] Part of the OKF

to create from the CC-licensed work. As already discussed, the database right does not restrict the production of non-database derived works (such as infographics) and so such things would not be "adapted material". This means that no share-alike or attribution obligations attach to that material.

In other words, ODC-ODBL takes a more aggressive approach to "share alike" than CC-BY-SA for non-database works derived from databases.

Applying a licence

Applying a licence should be relatively straightforward. From an intellectual property point of view, you do not need to prove that someone using your material has agreed to a licence. Rather if they are using it, they would have to show that they had a licence.

All that is needed is:

- reference to the licensing information in a clear place (eg the footer of a website)
- a clear statement of what is licensed and what (if anything) is not

Both CC and ODC licences have clear instructions on how to apply their licences.

Non-proprietary protection

Even where data is not "owned" or capable of ownership, the proprietor of a site making the data available may be able to make it difficult for others to use the data ways undesirable to the proprietor.

Computer Misuse

In the UK it is an offence[22] to access any computer in order to obtain access to data where:

- access to the data is unauthorised
- the offender knows that it is unauthorised

[22] Section 1, Computer Misuse Act 1990.

This is particularly relevant to those involved in "scraping" data – that is automatically simulating the action of a web browser and extracting data from the web pages read. This might, in appropriate circumstances, extend to accessing a website intending to use the data there in a way not authorised by the website owner.

The offence is moderately serious – leading to a maximum prison sentence of 2 years. It may also be committed at least partially across jurisdictional boundaries. Provided some part of the offence happens within the UK (eg if either the computer accessed, or the person attempting the access is UK based) then there is an offence within the UK.

This has implications if you use data from an online data source without the permission of the data provider. It may also be a problem if data you are supplied withhas been obtained in that way. As a practical matter, if you run a business relying on data provided by someone else you should strongly consider negotiating licence terms with them so you have some confidence their data source will continue to exist or be provided in a particular way.

The take away lesson for licensing of open data on a website or otherwise on the internet is that:

- your licence should be clear about what kind of access is allowed
- you can, in principle, prevent particular people using your data by expressly prohibiting it in your licence and making it clear that you consider any breach of those terms a criminal act.

Other use restrictions

The most significant other situation in which a restriction on use may affect the lawfulness of open data is where doing so is in breach of contract. Eg where someone signs up to terms and conditions on a website and then downloads data to use in a way forbidden by those terms and conditions.

That would be a breach of contract for which the website owner could sue. They might also be able to obtain an injunction (that is, a court order) to prevent the user from breaching the conditions of the contract.

Liability for data

Some forms of open data may create liability for the publisher. The subject of liability for online content is too broad to be covered in this course. It is possible, for example, for an open data set to contain information that amounts to a defamatory statement and thus risk defamation liability, itself a highly complex subject. Situations of that kind ought to be rare.

Personal Data protection

Open data may also be personal data. Personal data protection law is mostly harmonised across the EU by the Data Protection Directive.[23] In the UK personal data is governed by the Data Protection Act 1998. In the directive, "personal data" is defined as:

"any information relating to an identified or identifiable natural person"

"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 factors specific to his physical, physiological, mental, economic, cultural or social identity"

This means the following are likely to be personal data:

- email addresses
- patient data with the patient's name removed
- chatty blog posts saying things like "I met my friend Pete today"

Personal data law is too complex to deal with in a morning's talk. There is considerable guidance on the Information Commissioner's website (possibly too much) and elsewhere. The take-away for open data is that:

- processing personal data requires the satisfying of numerous legal requirements contained in the Data Protection Act 1998
- many of these can be avoided with the **informed** consent of the data subject
- suitably anonymising data – which may also require aggregation – will prevent it being personal data

Removing personal identifiers is not usually enough, because the natural person may still be identifiable. For example removing a patient's name from patient records but leaving an identification number that could be used to identify them would not prevent the data from being personal data. In practice anonymisation will usually require some form of aggregation of information or some other way to remove any practical possibility of "re-identifying" the data subject.

The Information Commissioner publishes a code of practice on anonymisation.[24] An Open Data Institute sponsored initiative the "UK Anonymisation Network"[25] can give advice on anonymisation.

Take away for licensing: in most cases ensure that your data does not contain personal data or that it is properly anonymised.

[23] Directive 95/46/EC

[24] http://ico.org.uk/for_organisations/data_protection/topic_guides/~/media/documents/library/Data_Protection/Practical_application/anonymisation_codev2.pdf

[25] http://www.ukanon.net/

Accuracy of data

Delegates to this course have expressed concern about their liability for publishing inaccurate data. If the data is supplied under a contract – for example where there is a paid-for API – then there may be a risk of liability if the contract does not properly make clear the problems that there may be with data quality. A clearly worded statement together with a disclaimer of liability should be sufficient.

It is also possible to be liable for inaccuracies in information under the common law tort of "negligent misstatement". If an inaccurate statement is made such that:

- someone – the recipient – relies on that statement for some particular purpose
- it was reasonable for them to rely on the statement for that purpose
- the maker of the statement intended the recipient to rely on the statement in that way
- the recipient suffers economic loss as a result of their reliance on the statement

Then the maker of the statement **may** be liable for the recipient's loss. The list of criteria is fairly tough. The courts will generally not think that it is fair to impose a liability unless the relationship of the parties is fairly close. In most of the reported cases the maker of the statement is found not to be liable.

For example, although auditors of a company's accounts may owe limited duties of care to the shareholders, they do not owe that duty to third-party investors who may be misled into investing as a result of an inaccurate statement.[26]

My view is that this is an unlikely form of liability to occur, but it may be avoided by attacking the "reasonable reliance" criterion. When publishing data that may contain inaccuracies (the almost invariable rule) a warning that the data may contain inaccuracies etc in much the same way as a disclaimer in a contract described above, should make it very unlikely for liability to be found.

---

[26] Caparo Industries v Dickman [1990] 2 AC 605