**Code of Practice on Datasets**

**General**

In its response to the Justice Committee’s report on the FOIA, the Government indicated its intention to introduce an exemption for pre-publication research, along the lines of the Scottish legislation. We are not aware of any timetable for the enactment of such an exemption, whereas it has been announced that the new dataset provisions in sections 102 and 103 of the Protection of Freedoms Act will be brought into effect in April 2013. If the dataset provisions come into effect before the new exemption, which seems likely, will datasets relating to pre-publication research be excluded from the commencement order that brings into force sections 102 and 103 of the Protections of Freedoms Act or otherwise protected in the period before the new exemption?

The code of practice may not be the appropriate place for this but it would seem necessary to consider how the dataset provisions interact with the action being taken by funders and the research community to develop open data and to make such data intelligible and reusable (see ‘Science as an open enterprise’, Royal Society 2012, especially in relation to the challenges and costs associated with ‘intelligent openness’).

**Details**

***iv. Standards applicable to public authorities in connection with the disclosure of a dataset***

Paragraph 16

The requirement to accompany the dataset with ‘*a sufficient amount of metadata and contextual information about how and why the dataset was compiled or created in order that users may fully comprehend the dataset’* seems to require the public authority:

1. to create information; whereas our understanding is that the FOIA is intended to require the disclosure of existing information but not the creation of information in response to a request; and
2. to second guess the nature and extent of information that will be required in order for users to ‘*fully comprehend*’ the dataset. What does ‘*fully comprehend’* mean? Can a public authority make any assumptions about the existing knowledge of the user? It is very likely that research datasets will not be capable of re-use by lay persons without substantial explanations being provided. In our view, the formulation ‘*fully comprehend’* is unreasonably broad, placing the onus on the public authority to make up an unquantifiable knowledge deficit on the part of the user.

The provision of contextual information is likely to be particularly onerous in relation to large datasets[[1]](#footnote-1), which have been compiled over many years and which may contain hundreds of millions of data points. Annotating such datasets to ensure that users ‘*fully comprehend*’ the dataset would probably be unmanageable and very costly.

The draft code states that the provision of contextual information is part of a public authority’s duty to provide advice and assistance, but such a requirement appears to go beyond the scope of section 16, as set out in the Code of Practice issued under section 45(2). Part II of that code suggests that the main purpose of section 16 is to require public authorities to provide information about procedures (so that the public can understand how to make a request) and to help applicants clarify the nature of the information being sought (so that the authority can identify and locate it).

The requirement to provide contextual information applies ‘*so long as there are no good reasons for it not to be provided’*. This will help, provided the ICO regards practical considerations of the sort described above as a sufficiently good reason.

We envisage that in its present form, we would find it difficult to comply with the requirement to provide contextual information more often than not. We suggest that it be amended along the following lines:

*‘Where reasonably practicable, published datasets should be accompanied by information about how and why the dataset was compiled or created.’*

**V. Giving permission for datasets to be re-used**

*Paragraphs 19 -21*

Ascertaining whether copyright and/or database rights are owned solely by the public authority will be potentially burdensome in relation to large research datasets, which have been collected over a number of years by multiple institutional and individual collaborators and funded by several external bodies. Much of the research undertaken by this University in the medical field involves datasets of this type: : some were initiated at a time when IP rights were not considered an issue for such datasets and so IP issues were not addressed in any agreements which may exist. The task will require specialist expertise, because of the complexity involved, and will need to be carried out with great care and attention to detail, given that the cost of getting it wrong could be a claim for substantial damages. It is unlikely that such a process could be completed within 20 working days but the FOIA only allows for extensions of this limit in relation to the application of the public interest test. Even if there is scope for streamlining procedures, as suggested in paragraph 21 of the draft code, it will still be necessary to consult any third party with a possible claim to IP rights, and it is this aspect of the process that is likely to be the most time-consuming and resource-intensive, particularly where there are a large number of parties involved and/or disagreements arise.

The time that a public authority should take to ascertain whether it is the sole owner of a relevant copyright work (and therefore in a position to comply with section 11A(2)), should be limited in some way. Section 12 of the FOIA will not apply, since searching for third party IP rights is not an activity that can be taken into account for the purpose of estimating whether the cost of a request exceeds the appropriate limit. We suggest therefore that the code be amended so that a public authority is required only to use reasonable endeavours or to take reasonable steps or to abide by some other similar formulation.

*Paragraph 22*

This paragraph states ‘*most datasets can be re-used without charge’* and that other [charging] licences will only be used in ‘*exceptional circumstances*’. It is unclear what circumstances can be taken into account in deciding whether a commercial licence may be offered and what constitutes an exceptional circumstance. The question might be put, should they be used without charge? In the case of this University, these datasets represent a charitable asset and as a charity, the University has an obligation to maximise the value it obtains from its assets for the purpose of contributing to achievement of its charitable objectives. Some datasets will have no particular commercial value, but it is going too far to say that this applies to ‘most’ datasets.

Given current financial conditions, it is particularly important that public authorities should be able to take advantage of new revenue opportunities. We suggest therefore that the draft code be amended so that public authorities are not unduly restricted and to provide specific recognition that it would be appropriate to use the Charged Licence in paragraph 25 where the dataset has commercial value to the public authority.

**vii. Considering publication of datasets as part of a publication scheme**

In addressing what ‘*appropriate’* means for the purpose of determining whether it is appropriate to publish a requested dataset, the draft code gives as examples cases where the data is exempt from disclosure or where there is ‘*unlikely to be a general public interest in the dataset*’. The first example is implicit in the provisions themselves – it would be illogical to publish a dataset under a publication scheme if it had already been refused in response to an individual request. The second example is helpful.

It would also be helpful if the draft code provided that a public authority could have regard to the resource implications of publication, in relation to both individual datasets and in general. Such provision should also be made in relation to paragraphs 33 and 34 of the draft code, which encourage public authorities to be proactive in publishing datasets.

The draft code suggests that authorities should maintain inventories of datasets (paragraphs 34 and 40) as a means of helping them manage the publication of datasets and as a source of assistance to the public. The creation and maintenance of an inventory of all datasets would be a substantial undertaking for this University and one of doubtful value if most are likely to be covered, at least for a significant part of their existence, by the new exemption for pre-publication research data.

1. E.g. the Million Women Study: <http://www.ceu.ox.ac.uk/research/31/> [↑](#footnote-ref-1)