Thoughts and reflections on the HEFCE “Policy for open access in the post-2014 Research Excellence Framework”

Authors: Petr Knoth and Zdenek Zdrahal

The following comments reflect our views on the *Policy for open access in the post-2014 Research Excellence Framework* further referred to just as the *HEFCE Policy*. These views are our own and do not necessarily reflect the official position of the Open University. We hope our comments will be useful for further improving and understanding the policy and its potential consequences.

Overall, we welcome the HEFCE Policy and believe it might be beneficial for the scholarly communication in the UK. However, we have noticed a few inconsistencies, which we would suggest to be clarified.

The concept of “open access” (OA) as used in the current version of the *HEFCE Policy* is not fully compatible with the Budapest Open Access Initiative definition of OA, i.e. the most authoritative definition of this concept.

> *“By open access to [peer-reviewed research literature], we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited.” [BOAI, 2002]*

According to the BOAI definition, open access is defined in terms of free access and reuse rights. While the HEFCE Policy supports the former, it still imposes a number of restrictions on the latter. In the context of the HEFCE Policy, open access refers rather to the free right to access and read research outputs, without necessarily having free reuse rights.

We will now comment on individual paragraphs as referenced in the HEFCE Policy document:

Paragraph 2. The word *anyone* should in our view refer to both humans and software.

Paragraphs 11 and 12. Paragraph 11 a. indicates that conference proceedings with ISSN are required to comply, while paragraph 12 states that they do not meet the definition. We suggest to clarify the wording.

Paragraph 17. We think arXiv.org might not be the best example as the majority of articles in arXiv.org are not OA. In addition, we think that this paragraph should require that the licence of the content must be clearly stated. This statement must be readable for both humans and machines. With respect to machine readable licences, repositories should support at least one of the existing standards, such as RIOXX/V4OA, OpenAIRE or NISO.

Paragraph 21. We welcome this requirement and believe this is a very important point.

Paragraph 25. We believe that suggesting CC-BY-NC-ND is not a good idea. This licence forbids many text-mining applications including basic search, in particular the use full-text snippets in search results, and multi-document summarisation. It also does not permit commercial use, implying that even the main search engines, such as Google Scholar and Microsoft Academic search, will be effectively breaching the law (due to both commercial use as well as the production of derivatives), unless they stop indexing this content.

Paragraph 34. We believe that the first sentence is in contradiction with paragraph 21. We would suggest that the policy requires repositories to allow the download and indexing of repository content by automated tools. This should be only subject to a fair user policy applying to all automated tools without exceptions. We believe that if this paragraph remains unchanged, it might be complicated to monitor the uptake and compliance with the HEFCE Policy.

Paragraph 37. We find it difficult to understand clause c) and are worried that the clause allows multiple interpretations. Some publishers currently do allow academics to put OA publications online, but disallow them from putting them in a repository or do not support OA completely. We are worried that the paragraph can be interpreted so that the author can use this clause as a justification for publishing with these publishers. This might be a potential “hole in the policy.”

Annex A.

Paragraph 2. Text mining is not a new practice. The history of natural language processing starts in 1950s [NLP]. Significant achievements in mining research outputs date back to 80s. An influential article called Undiscovered Public Knowledge [Swanson, 1986] which demonstrates the power of text-mining research outputs has already been published in 1986.

Paragraph 3. The paragraph states correctly that licensing and copyright are a limitation. In addition, an equally important limitation is caused by the lack of interoperability of systems and by the lack of fair access to data. More specifically, big players, such as Google, are typically granted unrestricted access to open access research outputs, while researchers and public (non-commercial and/or not-for-profit) services are denied free access. We think this issue should be considered.

In addition, this paragraph correctly states that text-mining might imply the creation of derivative works, however the licence proposed in Paragraph 25 does not allow many forms of text-mining, including basic search.

Paragraph 4. Text-mining software usually does not require a high number of requests as text-miners typically do not access documents individually, but rather through data dumps collected by harvesting systems. In addition, we believe that the allowed extent of *restricting bulk access and download by software* should be clarified. In our view, any restriction should be based on the technical limitations of repositories and should apply in the same form to everybody. Our experience indicates that these restrictions create significant barriers for researchers, but are mostly not applied to big commercial players. This results in an unfair advantage for some at the expense of others.

Paragraph 5. We believe that the use of Non-Commercial licences does not sufficiently support text-mining. On one hand, it will mean that existing companies including Google, Microsoft, Mendeley and ResearchGate will be breaching the copyright. As many of these organisations are economically strong, it is unlikely to expect they would change their practices as result of this policy. On the other hand, the risk of breaching the copyright will be a substantial barrier for innovative companies. These companies would be prevented from utilising text-mining to develop application benefiting the whole research sector.

References

[BOAI, 2002] http://www.budapestopenaccessinitiative.org/boai-10-recommendations

[NLP] http://en.wikipedia.org/wiki/Natural_language_processing

[Swanson, 1986]
http://www.jstor.org/discover/10.2307/4307965?uid=3738032&uid=2129&uid=2&uid=70&uid=4&sid=21103755973167