chem-bla-ics

Archiving, but not really

Published August 6, 2025

Citation
Willighagen, E. (2025, August 6). Archiving, but not really. *Chem-bla-ics*. https://doi.org/10.59350/vwd81-p8z85

Keywords
Publishing

Abstract
Mike Taylor wrote up a post about the various things a journal article is doing, the first being a scientific report. We put a lot of money in establishing a scientific track record. In the past 30 years how we publish our research and how we archive it has changed significantly. If you read my blog more often, you know I have been critical of the performance of many publishers.

Mike Taylor wrote up a post about the various things a journal article is doing, the first being *a scientific report*. We put a lot of money in establishing a scientific track record. In the past 30 years how we publish our research and how we archive it has changed significantly. If you read my blog more often, you know I have been critical of the performance of many publishers. Springer Nature was so disappointing that after 5 years I stepped down as Editor-in-Chief (of two) of the Journal of Cheminformatics. There is so much that must be done better.

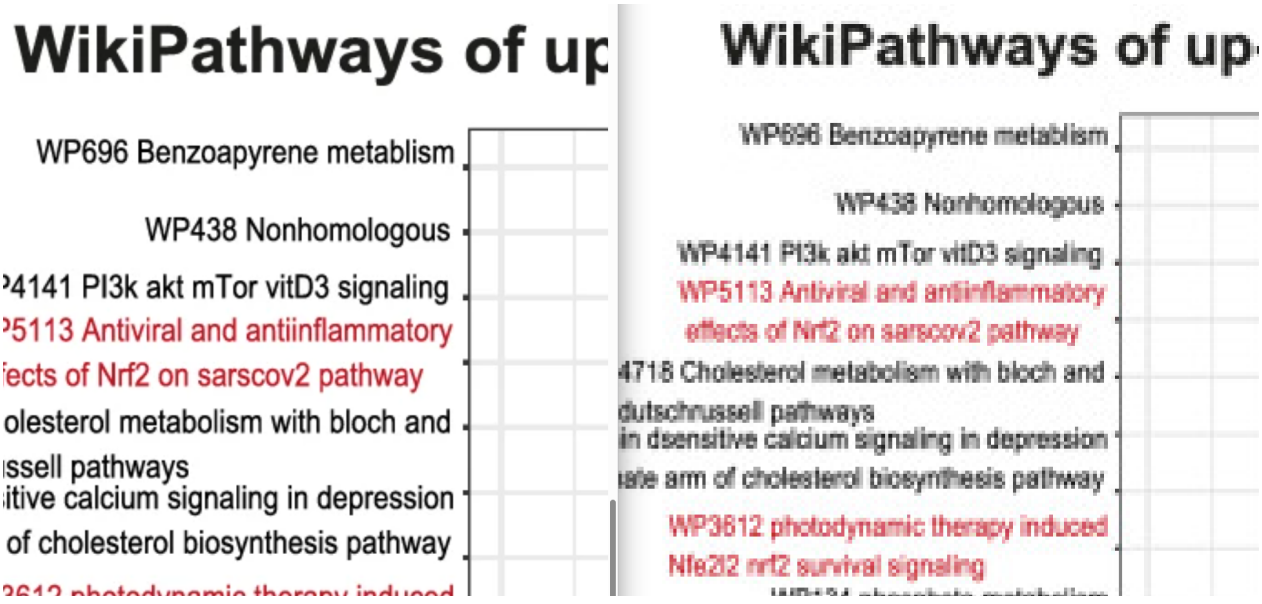
But in the most recent iteration, triggered by some work for WikiPathways, I was using Europe PMC to find articles that mention *WikiPathways* and then search in the full text for the string `WP`, as a trigger for the possible mention of WikiPathways pathway identifiers, which look like `WP4846`. The use of *compact (resource) identifiers* (see doi:10.1038/sdata.2018.29) is minimal, but at least some articles use identifiers.

That allows me to extend our WikiPathways knowledge graph of articles citing specific pathways. At the time of writing, we collected 2509 citations from 440 different articles to 883 different pathways. Now, I want to blog about that more, but it's related to an observation.

Information loss

Now, back in the late ninities I learned about GNU/Linux and after playing with Red Hat and Suse, I settled for Debian. One of the things I learned is that, generally, information corruption (like data loss) is an absolute red flag, a no-go, a total showstopper.

And then we have this in publishing, the one area where data corruption must also be a no-go:



In this image, the left side shows a screenshot of the publisher version of the article and on the right side the version in Pubmed Central (PMC). PMC has been an important project to archive full text versions of articles:

> *11.2 million articles are archived in PMC.*

chem-bla-ics

So, this is **really bad**! The archived version is not really useful. As a human I already struggle to read the degraded image, let alone an algorithm.

Does that matter? Yes, projects like the awesome Pathway Figure OCR (see doi:10.1186/s13059-020-02181-2) depend on images to be FAIR enough to extract information. (Side note: yes, these images should be vector graphics, but commercial publishers decided about twenty years ago that they could not care enough.)

At this moment, I do not know where the information is lost. Maybe PubMed Central is storing the images in a low resolution. Maybe the publisher provides PMC with a low resolution image. But to me, this must be solved as soon as possible. This is utterly unacceptable.

I wonder what the authors of the article (doi:10.1186/s13287-025-04166-z) I took as example think of this.