Why simple content labelling is guaranteed to fail

Phil Archer, writing in a personal capacity.

Between 2000 and 2008, I worked for an international membership organization called the Internet Content Rating Association (ICRA). A year after I was made redundant I wrote a lengthy paper on why that initiative failed, despite its political attractiveness and significant industry support. The purpose of this short paper is simply to offer a succinct version of that as a warning. A child protection solution based *solely* on metadata provided voluntarily by content creators that is then read by filters that act *purely* on the basis of that metadata will fail just as completely as ICRA and its predecessor, RSAC*i* did.

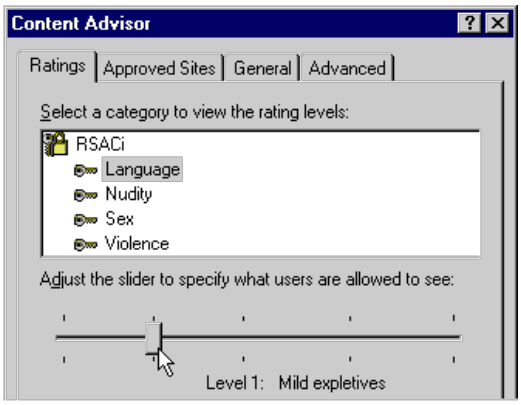
For some people, restrictions on access to content is nothing other than censorship. Like all absolutist positions, that is nonsense. Of course children should not see all manner of content that is available to adults. Preventing that was the simple aim behind those early "content rating" initiatives. But the threats are much greater; it shouldn't be possible to target children in grooming attacks or any number of other online dangers. The question is, how to achieve that protection reliably while respecting the openness that underpins the Web and that is coming under ever-more concerted attack, including from democratic countries.

It's not a new topic.

The first major piece of standardization work carried out at W3C was the creation of the Platform for Internet Content Selection (PICS). Just as today, there were arguments then about the balance between censorship on the one hand and the need for protection on the other, especially for children. The ideas behind it were well-thought through, technically sound and politically attractive:

- objective descriptions are provided by the content producer in a machine-readable format;
- software can read those descriptions and control access in accordance with the user's own views (or the views of their parents or employers).

Microsoft built support for this into Internet Explorer 3 as the "Content Advisor" and retained it through all its future versions. Netscape 4.5 included it too.



The interface for Internet Explorer's Content Advisor. Under the user's password control, the slider for each category could be adjusted to block or allow different types of content.

The idea of using metadata to distinguish types of content wasn't just applied to online safety. My role at ICRA led to me spending significant time as a proud member of the W3C Mobile Web Best Practices Working Group, not because of the issues around youngsters using their phones safely but because the plan was to declare websites that were mobile-friendly by applying a metadata scheme called Mobile OK. Heck, in effort to be modern and forward-looking, we even developed a whole new successor to PICS that used a combination of RDF, XML and GRDDL to achieve much the same thing. Ever heard of the Protocol for Web Description Resources (POWDER)? Let's assume not.

As I learned the hard way, it's not about the technology or the standards.

The Achilles Heel of the whole approach taken by ICRA was simple: what about content that isn't labelled? The choice is binary:

- accept all unlabelled content
- reject all unlabelled content

Pick one.

Since the overwhelming majority of online material will not be labelled, you're likely to accept it all – which means the whole thing is a waste of time.

Content filtering wasn't bad 20 years ago and of course it's even better today thanks to advances in AI. Nowadays it relies on a combination of lists of domain names and on-the-fly analysis. Metadata, including blocks of JSON-LD created using the schema.org vocabulary, could aid that on the fly analysis. Want to block hate speech? OK, look for bad grammar, short sentences and an over-use of bold and upper case letters. Want to block porn? It's really not that hard.

It's the scams, the coercion, the disinformation, the manipulative preying on the vulnerable that are the real dangers. That and the algorithms that promote only what gets more clicks without a single parameter to mitigate the harm that might be done.

No amount of free speech-supporting metadata and parental choice is going to address that.

Paper submitted to the IAB/W3C Workshop on Age-Based Restrictions on Content Access. This is a personal submission and should not be taken as the views of my current or any former employer.

30 July 2025

Back to diary