FREE-RANGE SPIDERBOTS!

LUC BORUTA THUNKEN THUNKEN, INC.

Reynold Brown, *Tarantula*, 1955.

FREE-RANGE WHAT!?

The **robots exclusion standard**, a.k.a. **robots.txt**, is used to give instructions as to which resources of a website can be scanned and crawled by bots.

Invalid or overzealous robots.txt files can lead to a loss of important data, breaking **archives**, **search engines**, and any app that **links or remixes scholarly data**.

WHY SHOULD I CARE?

You care about open access, don't you? This is about **open access for bots**, which fosters **open access for humans**.

MIND YOUR MANNERS

The standard is purely advisory, it relies on the **politeness** of the bots. Disallowing access to a page doesn't protect it: if it is referenced or linked to, it can be found.

We don't advocate the deletion of robots.txt files. They are a lightweight mechanism to convey crucial information, e.g. the location of sitemaps. **We want better robots.txt files.**

BOTS MUST BE ALLOWED TO ROAM THE SCHOLARLY WEB FREELY

Metadata harvesting protocols are great, but there is a lot of data, e.g. pricing, recommendations, that they do not capture, and, at the scale of the web, few content providers actually use these protocols.

The web is unstable: content drifts and servers crash, this is inevitable. Lots of copies keep stuff safe, and crawlers are essential in order to **maintain and analyze the permanent record of science**.

We want to start an informal open collective to lobby publishers, aggregators, and other stakeholders to **standardize and minimize their robots.txt files**, and other related directives like **noindex** tags.

OUR FIRST VICTORY

In September, we noticed that Hindawi prevented **polite bots** from accessing pages relating to **retracted articles** and **peer-review fraud**. Hindawi fixed their robots.txt after we brought the problem to their attention via Twitter. We can fix the web, one domain at a time!

ADMIT ONE

Interested? Drop us a line at contact@thunken.com