TDMRep & EPUB

EDRLab

EU CDSM Directive

- If they have legal access to resources on the Web, operators can freely download and use them for Text and Data Mining (TDM). This is an exception from copyright

- TDM is defined as "Any automated analytical technique aimed at **analysing large amounts of text and data** in digital form in order to **generate information which includes but is not limited to patterns, trends and correlations**".

- Publishers can opt-out from this exception using machine readable means.

- Opt-out is void if the purpose of TDM is scientific research.

W3C TDMRep CG

- AIE (Italy) and EDRLab created a TDM Reservation Protocol CG in 2021.
- 45 people joined, mostly from the publishing side.
- We used the W3C CG blog to post news (see references).
- The final report was released on Feb. 2022.
- A POC was developed by cairn.info and Seraphine.legal (both French)

TDMRep

- Objective = blocking a class of robots, not one specific robot. And being reeaally simple to implement.
- A boolean property: opt-out or not
- An optional url: if present, means "opt-out unless ...".
  - Points to an ODRL 2.2 json resource (templates available).
  - Contains the publisher's contact and conditions for obtaining mining rights.

How?

- Properties are named "tdm-reservation" and "tdm-policy"
- Three ways to express them, currently:
  - In the HTTP header of each resource
  - In a well-known file (tdmrep.json) hosted on the scraped website
  - As html metadata (if the resource is html)

Status of deployment

- New interest triggered by the growth of generative AI.
- The Federation of European Publishers (FEP) supports the solution. Geste (federation of online publishers, France) also.
- Ouest France (largest regional newspaper, France) has tagged its website. L'Équipe (sports newspaper), le Télégramme (regional newspaper), FranceTV (public broadcaster) seem to follow.
- New users from the publishing book sector in Sweden and the Netherlands. Federation of Screenwriters in Europe also interested.
- Urgent need: reach a consensus around the fact that "TDM" covers "AI".

New request: EPUB

- Some publishers (e.g. Actes Sud) would like to embed the properties in the EPUBs they publish.
- Nobody seems to care about embedding it in ONIX records (which would also make sense).

Proposal

- Use OPF metadata. Specify that the directives cover every resource in the EPUB.
- Define a new namespace with prefix "tdm" (or "tdmai"?)
- Simply add
  - tdm:reservation: boolean
  - tdm:policy: url
- The TDMRep CG can write the specification. ok?

References

- Work documents: https://w3c.github.io/tdm-reservation-protocol/
- TDMRep CG page: https://www.w3.org/community/tdmrep/
- TDMRep 2022 specification: https://www.w3.org/2022/tdmrep/