

De nouveaux modèles de plateformes reposant sur les technologies du web sémantique et l'automatisation des traitements ont fait leur apparition, à l'instar de la plateforme de publication science.ai qui propose des contenus scientifiques aux données riches et ré-exploitable. Décryptage par Robin Berjon, expert en technologies du web et membre actif des groupes de standardisation au sein du W3C.

Le Web, renaissance de la publication de recherche

La publication de recherche est dans un sombre état.

Si le volume d'articles publiés ne cesse de croître, les modalités de leur production semblent incapables de dépasser leurs archaïsmes. Ce volume pose lui-même problème, particulièrement dans les domaines les plus prolixes comme la recherche biomédicale, car la pauvreté des métadonnées et de l'indexation des articles publiés rend difficile voire impossible l'exploitation efficace du contenu existant. Il apparaît d'ailleurs qu'une écrasante majorité des citations pointe vers des publications récentes, suggérant qu'une part non négligeable du savoir produit est en tout état de cause oublié, et ce rapidement^[1]. Les formats employés pour la mise en ligne de ces publications font typiquement montre d'une qualité de production déplorable. Quand le PDF est employé, il ne fonctionne au mieux que très péniblement sur des terminaux mobiles et présente de sérieuses barrières d'accès pour certaines personnes en situation de handicap ; si c'est du HTML il est de piètre facture, dénué de sémantique, et d'ergonomie inexistante.

S'ajoute à ces éléments la question du coût : les frais de publication augmentent de façon constante, au point que certains laboratoires limitent les publications de leurs chercheurs, et les tarifs de consultation ont suffi à rendre « Sci-Hub », fournisseur de plus de 50 millions d'articles piratés, mondialement célèbre.

Ce constat n'est pas nouveau, ces problèmes sont connus de longue date. Pourtant, la situation semble n'avoir fait qu'empirer et plusieurs tentatives d'amélioration se sont soldées par un échec. Qu'est-ce qui permet de penser qu'un changement peut désormais survenir ?

DES MÉTADONNÉES VALORISÉES PAR LES TECHNOLOGIES DU WEB

Tout d'abord, la technologie est enfin mûre. L'avènement de HTML5 a rendu la plateforme Web suffisamment solide pour une publication de qualité, et la réussite du projet schema.org^[2], un système de métadonnées soutenu par les principaux moteurs de recherche du Web, a banalisé l'usage de l'enrichissement sémantique à plus de 10 millions de sites.

À titre d'exemple, le site du *New York Times* a récemment commencé à utiliser ces métadonnées pour ses recettes de cuisine. Celles-ci sont interprétées correctement par les moteurs de recherche généralistes, permettant ainsi aux utilisateurs de trouver bien plus facilement le gâteau au chocolat qui leur sied. Résultat : leur trafic a rapidement augmenté de plus de 50 %^[3].

Par ailleurs, le projet Sci-Hub et ses articles illégalement gratuits, dont il est établi que le monde entier l'utilise^[4], met les éditeurs sous pression. En effet, outre la mise à disposition gratuite du contenu il

[1] Della Briotta Parolo, Pietro, Raj Kumar Pan, Rumi Ghosh, Bernardo A. Huberman, Kimmo Kaski, and Santo Fortunato. 2015. "Attention decay in science." <http://arxiv.org/abs/1503.01881>

[2] <http://schema.org/>

[3] New York Times. 2014. "The Full New York Times Innovation Report." *New York Times*. <https://www.scribd.com/doc/224608514/The-Full-New-York-Times-Innovation-Report>

[4] Bohannon, John. 2016. "Who's downloading pirated papers? Everyone." *Science*. 28 April. <http://www.sciencemag.org/news/2016/04/whos-downloading-pirated-papers-everyone>



Page d'accueil de la plateforme Pirate Sci-Hub

apparaît qu'un nombre croissant de chercheurs ont recouru à Sci-Hub même quand ils ont légalement accès aux articles, tout simplement parce qu'ils trouvent l'interface plus ergonomique que celle fournie par les éditeurs.

Sci-Hub n'est pas la solution, cependant. Outre son illégalité, c'est pour l'essentiel une source de documents sous format PDF, pauvres en sémantique et largement inaccessibles. Mais c'est un clairon qui réveille les acteurs de l'édition de recherche.

Le moment est donc propice au changement, et c'est précisément l'ensemble de ces problèmes que la plateforme de publication science.ai⁵ s'applique à résoudre.

SCIENCE.AI : UNE CHAÎNE DE PUBLICATION OPTIMISÉE

L'approche choisie est résolument centrée sur les technologies du Web moderne et sur l'automatisation du traitement. En termes de coûts, celle-ci permet de tirer les frais de publication tellement bas que le débat sur le libre accès devient caduc : il ne subsiste tout simplement plus de raison financière de faire payer la consultation.

Ceci se fait sans compromission au niveau de la qualité de production, le Web moderne permettant précisément de bien meilleures publications et utilisations que précédemment. Dès l'obtention du

trouvent que le Web ne présente pas le côté pratique du PDF, simple à télécharger. C'est sans compter sur les nouvelles fonctionnalités hors-ligne du Web et sur la convergence entre le Web et les formats de livres électroniques. Il est aujourd'hui possible d'utiliser du contenu Web de façon locale et déconnectée.

Finalement, comment obtient-on des métadonnées pertinentes ? Les auteurs sont en effet souvent peu coopératifs lorsqu'il s'agit de remplir de longs formulaires décrivant leurs travaux... Ce problème est résolu par le truchement de deux méthodes. D'une part, des techniques issues de l'intelligence artificielle fournissent là où c'est possible une classification du contenu dont la précision est comparable à celle obtenue par des intervenants humains. D'autre part, l'interface de mise en ligne du manuscrit met l'enrichissement naturel de ce document par les métadonnées dans un contexte compréhensible, ce qui encourage subrepticement les auteurs à fournir plus d'informations qu'ils ne le feraient face à un formulaire rebutant.

Si ce système doit bien entendu faire ses preuves, un partenariat commercial avec John Wiley & Sons, un des plus grands éditeurs de recherche, est déjà en cours. Afin d'en juger par soi-même, chacun peut s'inscrire à un système de démonstration, limité mais fonctionnel, sur le site <https://science.ai/>. Cette plateforme est conçue pour fonctionner à plusieurs

Une renaissance de la publication de recherche, par l'entremise des technologies ouvertes qui sont aujourd'hui disponibles, est à portée de main.

manuscrit, le contenu est converti en un ensemble utilisant à la fois le format de données sémantiques RDF et le format de texte structuré HTML. Ceci permet de publier un contenu non seulement riche en métadonnées (lesquelles sont utilisables par le projet « science.ai » lui-même comme par toute personne tierce car celles-ci sont imbriquées directement dans la page) mais aussi pleinement accessible indépendamment des terminaux ou des handicaps, le tout sur la base de standards ouverts. Bien entendu, cette transition connaît ses obstacles. Premièrement, les manuscrits fournis par les auteurs restent dans des formats dénués de sémantique et péniblement exploitables comme Word ou LaTeX. À cet effet nous avons établi un guide d'utilisation de Word (outil de plus de 85 % des manuscrits) qui, sans installation d'outil ou de feuille de style spécifique, permet d'extraire toutes les informations essentielles d'un document. C'est le "DOCX Standard Scientific Style" (ou DS3), disponible librement en ligne à <http://docx.science.ai/>.

Un autre obstacle est que beaucoup de chercheurs

échelles. À une extrémité du spectre un éditeur de grande taille peut s'en servir pour gérer de nombreuses publications, mais l'outil reste tout à fait accessible à une petite société savante, voire à un auteur seul, qui désire mettre en place une

revue. Un de ses objectifs est d'ailleurs d'offrir une interface suffisamment simple et accessible pour libérer l'innovation en matière de ligne éditoriale, de types de contenus, ou d'organisation de comité de lecture. C'est une plateforme permettant au plus grand nombre d'accéder potentiellement à la publication de recherche.

Une renaissance de la publication de recherche, par l'entremise des technologies ouvertes qui sont aujourd'hui disponibles, est à portée de main. Car si le bon sens est la chose du monde la mieux partagée, il s'agirait que la connaissance le soit aussi. Des articles apportant plus de précisions techniques sur certains aspects sont disponibles sur <https://research.science.ai/> (en anglais), et bien entendu nous attendons vos retours sur notre système lorsque vous l'aurez essayé !

ROBIN BERJON

Directeur technique,
science.ai – New York, USA
robin@berjon.com

[5] <https://science.ai/>