Partager le travail scientifique à l'âge numérique.

Pascal Guitton, Marie-Hélène Comte, Thierry Viéville

Partager le travail scientifique à l'âge numérique.

Contribution Inria.

Introduction

Ce texte s'intéresse au changement profond qu'a connu la publication scientifique sur une génération de chercheurs : encore tapée à la machine au début des années 80, elle est progressivement devenue uniquement numérique grâce à l'arrivée des systèmes de traitement de texte et d'édition comme LaTex ou Word par exemple. Cette apparition du document numérique s'est accompagnée d'une mise en ligne progressive tant des revues et des actes de conférence que des rapports de recherche. Par ailleurs, la nature même du document a évolué : il est devenu hypertexte, multimédia (son, vidéo) ou logiciel exécutable sans oublier les données d'expérimentations et les cours en ligne qui deviennent à leur tour publications.

Ces évolutions importantes ont conduit à une profonde modification de la fréquentation des sources de documentation par les scientifiques. Les bibliothèques étaient il y a 30 ans un des principaux accès à la connaissance alors qu'aujourd'hui, les chercheurs utilisent principalement moteurs de recherche et portails spécialisés pour chercher, accéder et lire les articles qui leur sont nécessaires. Ce constat ne signifie pas que les bibliothèques en tant que lieu ou centre de ressources soient devenues inutiles ou obsolètesmais leur utilisation a profondément changé et il est donc indispensable d'adapter leur rôle à cette réalité.

La publication scientifique numérique

Les scientifiques ont toujours communiqué le résultat de leurs recherches à leurs pairs de façon orale et écrite. La forme a évolué, du manuscript à l'imprimerie puis au traitement informatique de texte. Cette évolution n'est pas terminée et on constate aujourd'hui une diversification très importante des formats utilisés qui transforment les textes d'autrefois [1] en documents dynamiques incluant des liens hypertextes, des sons [2], des vidéos et même des codes exécutables [3, 4, 5]. Des réflexions sont actuellement menées pour publier des logiciels dans des revues adaptées en les accompagnant de données et surtout des paramètres d'environnement matériel qui autoriseraient leur exécution "un certain temps après" par exemple pour établir des comparaisons (notion de reproductibilité des expériences) [6].

Par ailleurs,face à l'explosion de la quantité et mutation de la nature des résultats scientifiques, la documentarisation de ces documents (méta-données documentaires et sémantiques, choix de mot-clés dans des vocabulaires contrôlés, …) devient un enjeu majeur [9]. En effet, des outils de veille associant des ressources humaines et algorithmiques sont devenus incontournables pour se donner les moyens de manipuler ces sources de connaissances à des échelles où elles ne peuvent plus être appréhendées par le seul cerveau humain [10].

La folksonomie, c'est à dire l'indexation libre et collaborative faite par les lecteurs et co-contributeurs du document, conduit à une autre dimension: celle de pourvoir enrichir le document de données sur les auteurs et les lecteurs de ces contenus [20]. C'est aussi un levier majeur pour aider à manipuler de manière pertinente ces grands ensembles de publications [11], [12]. D'une lecture solitaire retraduite dans la section bibliographique d'une thèse ou d'un article, nous passons à une lecture réactive et collective où le document s'enrichit au fur et à mesure de sa diffusion.

Les initiatives de type open-data permettent de fournir un socle à cette nécessité de partager les résultats scientifiques sous des formes ouvertes (à commencer par l'exemple historique [13]) mais bien au delà de versions numériques de documents papiers [14]. On parle ici de bases de données de benchmarking (par exemple dans des domaines divers [15,16]), de mise à disposition d'outils logiciels de comparaison de modèles, etc.

La publication scientifique vers de nouveaux publics

Par ailleurs, de plus en plus de scientifiques présentent également une partie de leurs résultats au grand public de façon à l'aider à mieux comprendre les bouleversements que connaît notre monde. Dans ce cas, on retrouve la même démarche que précédemment mais en adaptant la formulation à un public non spécialiste et

qui ne dispose pas des compétences des experts du domaine. Là aussi, document textuels, multi-médias et grains logiciels [7, 8] cohabitent.

Il faut comprendre que cette ouverture de la publication scientifique vers la création de contenus de médiation scientifique ne se limite pas à des articles dans des revues de curieux de science ou à un désir de médiatisation. C'est aussi dû, tout spécialement en science du numérique, à ce que la recherche scientifique est devenue organiquement inter-disciplinaire donc que les publications ne s'adressent moins souvent uniquement aux scientifiques spécialistes du domaine. De plus, la volonté de mettre les applications issues de la recherche au service des grands enjeux sociétaux actuels conduit à utiliser les publications scientifiques comme un outil de transfert vers le monde socio-économique. La forme éditoriale du document doit intégrer cet aspect multi-cible.

Il faut ausi comprendre que disposer de publications enrichies au niveau multi-média ou disposant de contenus interactifs de démonstration des résultats n'est pas qu'un enrichissement de forme. Ce sont les résultats scientifiques *eux mêmes* qui ont besoin de ces supports de présentation. Citons quelques exemples. En mathématiques, la partie calculatoire (au sens de calcul symbolique) de la démonstration formelle d'un théorème n'est plus simplement rédigée en langage naturel mais sous forme d'un code logiciel (par exemple en COQ [17]). Publier une démonstration devant ses pairs nécessite donc de publier le code et l'environnement pour exécuter le code. En neuroscience, il est devenu impossible de décrire de manière verbale ou sous forme de tableaux de chiffres un modèle de connectivité cérébrale dynamique sophistiqué ; l'objet scientifique à publier peut-être intrinsèquement un objet 3D (une représentation anatomique enrichie de données par exemple) dont chaque voxel contient des informations qualitatives et quantitatives. En sciences sociales, l'étude d'un réseau social ne peut plus se réduire à un narratif, mais se présente sous forme de visualisations scientifiques extrêmement sophistiquées, qui vont être réutilisées par des statisticiens ou des spécialistes de la théorie des graphes. On voit donc sur ces exemples caractéristiques que l'inter-disciplinarité conduit à partager les données et résultats scientifiques sous les formes enrichies décrites ici. On voit aussi que lors de la publication de résultats, on ne peut plus séparer les données du document qui les expliquent.

Il faut également noter que beaucoup de scientifiques ne sont pas uniquement chercheurs mais également enseignants et assurent des formations auprès d'étudiants (en formation initiale), voire de professionnels (en formation permanente). Il s'agit de dispenser des savoirs tant d'initiation ou de sensibilisation que issus de la recherche (notion de Formation par la recherche) à un large éventail d'apprenants. Au cours du temps, les formes écrites (livres, polycopiés, hyper-documents numériques) et pédagogiques (cours magistraux en amphi, cours en ligne de type Formation en Ligne Ouverte à Tout-e-s (FLOT ou MOOC)) ont également fortement évolué et aujourd'hui un FLOT peut être considéré comme une publication scientifique.

En marche vers des epi-journaux

Au final, il convient de s'interroger sur l'éditorialisation des documents scientifiques si l'on souhaite construire une alternative de qualité - et par conséquent crédible - aux revues commerciales dont le modèle économique impliquait de renoncer aux droits sur les contenus publiés. Si renoncer à nos droits sur les publications d'articles scientifiques textuelles était discutable, pour des contenus qui vont inclure nos données, c'est justement totalement inenvisageable. La notion d'epi-journal a donc vu le jour ; il s'agit de construire « au dessus » d'une archive ouverte des structures de type journal ou actes. La démarche est tout à fait similaire à celle d'un éditeur scientifique : diffusion d'un appel à communications, dépôt des propositions sur un site dédié en respectant une charte graphique, expertise par un comité de lecture dont la composition est publique, annonce des résultats aux auteurs, mise en ligne des articles retenus après réalisation des corrections demandées, référencement par les moteurs de recherche après saisie des méta-données associées. Voir par exemple, la conférence IHM'13.

Perspectives : quels axes de travail pour la suite ?

Nous pouvons au delà du constat de ce texte, expliciter les enjeux à court terme liés à ces sujets,

La recherche sur ces nouvelles formes de documents est encore à mener, à la fois au niveau de la visualisation scientifique que de la formalisation sémantique des faits scientifiques à partager.

Le développement des outils et systèmes nécessaires (exemple epi-sciences pour epi journaux) en est encore au stade expérimental. Il y a là à la fois un enjeu en terme d'interaction humaine avec ces données complexes et un enjeu de plateforme et de mise à disposition des logiciels appropriés.

Au delà des outils, la normalisation et standardisation des méta-données est en cours (par exemple les LOM pour les contenus à vocation pédagogique) mais l'adoption de ces normes et leur déclinaison à divers champs disciplinaires et un chantier ouvert

La sensibilisation et formation des scientifiques à ces nouvelles méthodes est un enjeu majeur et une urgence au niveau de la formation à la recherche et de la formation permamente de chercheurs. Cela inclut l'adaptation (structuration, formations) des professionnels IST à ces nouveaux métiers.

Et c'est justement le chemin que nous empruntons côté Inria, de manière partenariale avec toutes les structures de recherche avec qui nous collaborons au quotidien.

Annexe: quelques éléments clés du document 3.0

Très brièvement, essayons de donner une description précise de ce document scientifique dit, 3.0. Au delà du document papier 1.0 ou numérique 2.0 le document 3.0 (voir[1] pour plus de détails) est :

1. Documentarisé : donc doté de méta-données sémantiques du web 3.0 pour accéder à son contenu en terme de données et pas uniquement de texte ;

2. Hyper-texte : donc décomposé en grains reliés par des liens, de façon à accéder de manière modulaire à chacun de ces éléments, y compris multimédia (donc enrichis de graphiques, sons, animations, 3D).

3. Interactif : complété de portions de code pour interagir avec le contenu et se l'approprier pas uniquement par la lecture, l'écoute audio ou en visionnant une vidéo ou animation 3D mais en *manipulant* certains de ces éléments.

4. Participatif : intrinsèquement co-auteurisé, donc recevant des contributions des auteurs et aussi des utilisateurs du contenu, avec une ligne éditoriale bien cernée.

Bref : rien de "nouveau" juste l'aboutissement d'une évolution que l'on rend explicite ici.

Par documentarisé, on entend relié aux méta-données sémantiques du Web 3.0, c'est à dire l'atomisation des connaissances sous forme de «RDF»: *sujet propriété objet* ainsi que l'organisation en corpus contextuel standardisé, par exemple avec la LOM qui décrit les "objets pedagogiques" [21]. On obtient ainsi une mécanisation maximale des connaissances qui deviennent donc exploitables par les logiciels. Par exemple wikipedia devient DBpedia ce qui donne accès aux données contenues dans le texte. Il y a là un double enjeu de visibilité + ré-utilisabilité de tels documents.

Par hyper-texte on fait bien évidemment référence aux contenus du Web documentaire, décomposés en grains de modularité maximale, pour aider à la réutilisation maximale des contenus existants (ex. : on reprendra plus facilement une figure complètement synthétique facilement accessible qu'un long texte . .) déposés sur des plateformes idoines (ex: wikipedia, canalU, «SIL:O») et accessibles via un permalien. Il y a aussi un renversement de lieux sur le Web : au lieu de créer sa (encore une!) plateforme, on donne la possibilité que des parties de ces contenus puissent exister chez les autres, en syndicant les contenus sur des plateformes cibles. La notion de document devient un «arbre de grains» structuré en graphe ou «parcours» (ex. : découverte selon le public, ou selon le pré-requis ou l'acquis pour un contenu didactique). Ici l'index

[1] How the next generation of dynamic documents will improve the way we work

est “aussi important” que le texte. Un graphisme/son/vidéo devient un «arbre de plans» accompagné de son scénario détaillé. Il y a là un double enjeu d'interopérabilité + et de ré-utilisabilité.

L'introduction massive d'éléments interactifs, basée sur la notion d'«application» (grain logiciel), permet que cette notion de grains interactifs soit mieux formalisée et compréhensible du public cible. On identifie rapidement des formes d'interactions paramétrables, pour manipuler les données. Ce volet est encore largement à l'état de friche alors qu'il est bien mieux spécifié pour des documents à vocation pédagogique ou culturel [18].

Par travail participatif, on entend plusieurs aspects. D'abord pour ces grains de documents scientifiques, la collaboration entre un scientifique et d'un professionnel de l'IST, car une part de plus en plus grande de ces documents doivent tenir compte de l'existant, donc nécessitent un travail documentaire poussé (veille documentaire, identification des sources, collecte d'extraits pertinents, …). Puis la validation des sources est un élément primordial, et il faut tracer qui a écrit/validé quoi pour comprendre de quel contenu il s'agit. On doit ensuite chercher d'abord à contribuer à l'existant (ex. : contribuer à wikipedia puis en reprendre le contenu plutôt que de créer des ¨sous¨-plateformes) et rechercher à créer les grains manquants. Créer «une présentation» revient alors à «ré-assembler» les grains en corrigeant/validant complétant/améliorant les existants. Il faut alors accepter que la notion de «droit d'auteur» se dilue, et que les collègues IST deviennent des acteurs directs des publicatioins scientifiques.

Références

[1] Le premier article scientifique de l'histoire https://interstices.info/jcms/int_69786/le-premier-article-scientifique-de-lhistoire-de-linformatique

[2] Audio-slides, Elsevier (cliquer sur le bouton à droite de l’écran pour écouter un commentaire accompagnant la lecture de l’article) http://www.sciencedirect.com/science/article/pii/S1877343512001042

[3] Article of the future, Elsevier, cliquer sur la vidéo (5’30) http://www.articleofthefuture.com

[4] Executable paper, Elsevier, cliquer sur la vidéo (2’48) http://www.elsevier.com/physical-sciences/computer-science/executable-papers

[5] Enhanced publication, OpenAire http://www.openaire.eu/en/component/content/article/76-highlights/344-a-short-introduction-to-enhanced-publications

[6] Code repository, G. Fursin http://c-mind.org/repo

[7] Exemple d’applet, Interstices https://interstices.info/jcms/c_24839/jouez-avec-les-diagrammes-de-voronoi

[8] Exemple d’applet en chimie http://chem-file.sourceforge.net/data/carboxylic_acids/D-tartaric_acid_fr.html

[9] M**étadonnées et valorisation de l'information** http://bbf.enssib.fr/consulter/bbf-2006-04-0094-012

[10] Graphes RDF et leur Manipulation pour la Gestion de Connaissances. http://www-sop.inria.fr/members/Fabien.Gandon/docs/HDR_Fabien_Gandon.html

[11] A free service for managing and discovering scholarly references http://www.citeulike.org

[12] Making it easy to present your publications and share them with others http://www.researchgate.net

[13] An e-print service in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance and statistics. http://arxiv.org

[14] Open science data is a type of Open data focused on publishing observations and results of scientific activities available for anyone to analyze and reuse. http://en.wikipedia.org/wiki/Open_science_data

[15] Computer vision test images http://www.cs.cmu.edu/~cil/v-images.html

[16] The UC Irvine Machine Learning Repository http://archive.ics.uci.edu/ml

[17] A formal proof management system http://coq.inria.fr

[19] Comment utiliser le 3.0 pour que notre MINF soit ubiquitaire, participatif et attractif ? http://hal.inria.fr/hal-00756476

[20] Olivier Le Deuff. Du tag au like. La pratique des folksonomies pour améliorer ses méthodes d'organisation de l'information. Fyp éditions, 2012

[21] Présentation des standards : (LOM) – Learning Object Metadata, Rosa María Gómez de Regil, Doc'Insa, Lyon (2004) L'indexation des ressources pédagogiques, 2004, enssib, Villeurbanne.

Contributeurs

Pascal Guitton, Marie-Hélène Comte, Thierry Viéville.