
OpenArticles : libérez votre savoir !

Sujet : mise en place d'un service libre et gratuit
de dépôt et d'accès aux articles scientifiques



David LARLET

16 janvier 2006

Sommaire

1	État des lieux	2
1.1	Mécanisme de publication traditionnel	2
1.2	Ressources	2
1.3	Problèmes posés	2
1.4	Initiatives actuelles	3
2	Projet de recherche	4
3	Mise en œuvre	5
3.1	Point de vue humain	5
3.2	Point de vue technique	5
3.2.1	Architecture globale	5
3.2.2	Comptes utilisateurs	6
3.2.3	Attribution d'un score aux articles	6
3.2.4	Fonction de recherche	7
3.3	Point de vue financier	7
4	Perspectives	8
4.1	Auto-financement	8
4.2	Ouverture du code	8
4.3	Extension des services liés aux comptes	8
4.4	Internationalisation	9
5	Conclusion	10

1 État des lieux

La publication est une composante indispensable du processus de recherche. Elle permet à la fois de partager ses connaissances et d'obtenir la reconnaissance de ses pairs.

1.1 Mécanisme de publication traditionnel

Actuellement, le mécanisme de publication se déroule selon trois étapes :

- L'auteur de l'article envoie celui-ci à un éditeur de revues qu'il détermine en fonction de son audience et de son prestige. L'auteur doit parfois payer en fonction du nombre de page et de figures en couleur.
- L'éditeur envoie cet article à 2 relecteurs ou davantage (variable selon les revues) qui estiment la qualité de l'article et peuvent demander des précisions supplémentaires. Les relecteurs ne sont pas payés.
- En fonction de l'appréciation des relecteurs, l'article est approuvé par l'éditeur qui l'imprime et le vend.

1.2 Ressources

Plusieurs solutions s'offrent au chercheur de manière à se tenir informé des récentes découvertes scientifiques :

- L'abonnement aux revues papier.
- L'abonnement aux revues en ligne.
- La recherche en ligne des articles qu'il juge intéressants et leur paiement à l'unité.
- La lecture de journaux gratuits (le plus souvent numériques).

1.3 Problèmes posés

Le système actuel soulève plusieurs problèmes :

- Le libre accès à la science se trouve compromis. De plus, au delà de cet aspect idéologique, une importante proportion de découvertes scientifiques décrites au sein des articles scientifiques sont issues de la recherche académique qui est financée par les impôts des citoyens, il serait donc logique que cette connaissance leur soit accessible.
- Ce système est aussi à l'origine d'un coût non négligeable. On parle aujourd'hui du « triple coût » de la recherche pour l'État :
 - Financement du projet de recherche.

- Paiement des salariés qui relisent les publications sans rémunération.
- Financement des bibliothèques pour l’abonnement aux différentes revues.
- Enfin, ce système engendre une perte de temps considérable, que ce soit au niveau des délais découverte → publication ou au niveau des multiples authentifications pour accéder à des articles n’appartenant pas aux mêmes revues. De plus, les articles n’étant pas disponibles dans leur intégrité il est difficile d’implémenter des recherches d’articles s’avérant pertinentes.

1.4 Initiatives actuelles

De façon à faire évoluer cette situation certaines initiatives ont récemment vu le jour :

- En France, le CNRS, l’Inserm, l’INRA et l’INRIA ont mis en ligne des archives institutionnelles depuis le 22 mars dernier¹ conformément à leur engagement lors des conférences de Berlin² en octobre 2003 et de Southampton³ en mars 2005.
- En Angleterre, le rapport du Parlement du 20 juillet 2004 a recommandé que tous les établissements d’enseignement supérieurs du Royaume-Uni mettent en place des archives institutionnelles pour conserver les résultats publiés et les rendre consultables en ligne gratuitement⁴.
- Aux États-Unis, le NIH a présenté sa politique en faveur du libre accès à l’information scientifique le 3 février dernier, ce qui est déjà mis en application par certaines universités⁵.
- En Amérique du Sud, le projet SciELO (Scientific Electronic Library Online) est développé depuis 1997 dans le but d’améliorer la visibilité et l’accès aux résultats de la recherche scientifique sud-américaine⁶.

D’autre part, depuis 1999 existent des outils développés par l’Open Archive Initiative⁷ basés sur des formats ouverts garants de la pérennité des archives qui ont permis par exemple la création de sites comme arXiv⁸ constituant des archives ouvertes de pré-publications et de post-publications d’articles.

¹<http://www2.cnrs.fr/presse/communique/640.htm>.

²<http://www.zim.mpg.de/openaccess-berlin>.

³<http://www.eprints.org/events/berlin3/>.

⁴<http://www.publications.parliament.uk/pa/cm200304/cmselect/cmsctech/399/39902.htm>.

⁵<http://oaister.umdl.umich.edu/o/oaister/>.

⁶<http://www.scielo.org>.

⁷<http://www.openarchives.org>.

⁸<http://fr.arxiv.org>.

2 Projet de recherche

Historiquement, les éditeurs étaient nécessaires pour faire le lien entre ceux qui produisent des informations et ceux qui en cherchent. Avec l'avènement de l'Internet et de la numérisation de ces informations, la chaîne de distribution peut être de longueur nulle. Les éditeurs qui se sont maintenus grâce à un certain prestige voient leur fonction sociale se réduire d'année en année. Les éditeurs vendaient des revues, ces supports physiques disparaissant, le monopôle de ces éditeurs est en train de s'effondrer. C'est une évolution logique et les éditeurs n'ayant pas évolué en ce sens sont contraints à une disparition prochaine.

On peut aisément effectuer le parallèle de la situation actuelle vis à vis des publications et des éditeurs avec celle de la musique et des majors du disque. On assiste à une véritable révolution numérique dans ce domaine avec la création de nombreux sites¹ qui diffusent gratuitement de la musique libre².

Ce projet de recherche va consister à mettre en place un service libre et gratuit de dépôt et d'accès aux articles scientifiques.

Dans ce but, il va être nécessaire, dans un premier temps, de réaliser un site internet permettant :

- La consultation des articles scientifiques et la possibilité d'effectuer des recherches sur leur intégralité (contrairement à Pudmed³ qui n'utilise que l'abstract des articles).
- La notation des articles au moyen d'une fonction de score adaptée de façon à maintenir la qualité actuelle des articles publiés.
- La création de profils utilisateurs permettant de définir les relecteurs potentiels et de commenter/noter les articles lus.

Il est ensuite envisagé d'ajouter divers services spécifiques permettant le travail collaboratif ou la constitution de ressources communes (agendas par discipline, ...) par exemple.

¹<http://irate.sourceforge.net/> ou <http://www.jamendo.com> par exemple.

²<http://www.musique-libre.org>.

³<http://www.ncbi.nlm.nih.gov/entrez/>.

3 Mise en œuvre

3.1 Point de vue humain

Le public visé est clairement scientifique (chercheurs et étudiants principalement). À ce titre, il est nécessaire que celui-ci soit non commercial et qu'il soit dans la mesure du possible indépendant de son pays d'hébergement. L'établissement d'une charte propre au site et à ses engagements permettrait de clarifier cette situation. Ce projet doit recevoir l'approbation de l'ensemble de la communauté scientifique et des organismes de recherches associés.

Dans un premier temps, un énorme investissement des ressources va être consacré à la publicité du service. En effet, celui-ci ne pouvant être connu par la voie habituelle – à savoir la publication dans une revue – il va falloir utiliser tous les moyens possibles (directives gouvernementales, listes de diffusion, sites « partenaires », . . .) pour faire connaître le service.

Enfin, il est inutile de faire de la publicité si le service proposé n'est pas adapté et stable et donc à l'origine de l'adhésion des visiteurs. Il est nécessaire dans ce but d'opter dès le départ pour une solution pérenne et simple.

3.2 Point de vue technique

3.2.1 Architecture globale

Ce service est basé sur la participation active des visiteurs, que ce soit au niveau du dépôt des articles qui devra être fait volontairement par les auteurs ou au niveau de la notation des articles.

Étant illusoire de vouloir stocker l'ensemble des articles sur un serveur ou dans une base de données (on estime qu'1,2 millions d'articles sont publiés chaque année), la solution est le stockage des publications localement à l'échelle universitaire sur différents serveurs. Celui-ci sera réalisé au format XML conformément aux directives plébiscitées par le mouvement Open Archive Initiative¹ permettant un export après traitement au format HTML ou PDF. C'est la méthode actuellement retenue par eprints².

Grâce à ce système (cf. schéma récapitulatif 3.1, page 6), les services proposés vont directement récupérer le contenu des articles sur le serveur distant et seuls les liens vers ces ressources sont stockés sur le site. Les

¹<http://www.openarchives.org>.

²<http://www.osti.gov/eprints/about.html>.

problèmes soulevés concernant la rapidité de la recherche basée sur une telle architecture sont évoqués ci-après (paragraphe 3.2.4, page 7).

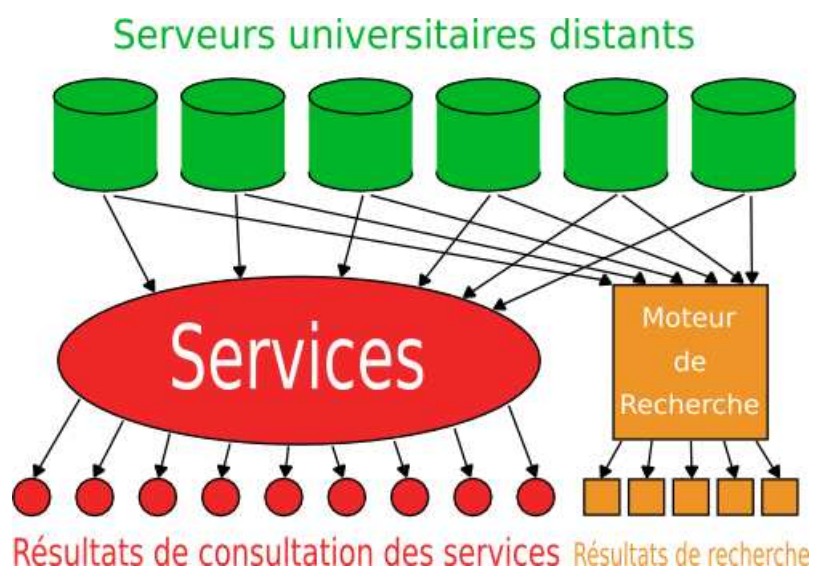


FIG. 3.1 – Schéma de l'architecture globale du service.

3.2.2 Comptes utilisateurs

- Ils doivent permettre de connaître, en plus des informations usuelles :
- Le domaine de recherche précis de façon à estimer la capacité du chercheur à être un relecteur ou non pour un article donné.
 - Le laboratoire de recherche pour permettre de localiser l'utilisateur et lui proposer des services associés (conférences proches, ...) mais aussi pour éviter que toutes les personnes d'un même laboratoire ne votent pour un article issu de ce même laboratoire.

3.2.3 Attribution d'un score aux articles

C'est la fonction clé du service principal. En effet, les revues sont aujourd'hui associées à un « impact factor » qui est en théorie proportionnel à la qualité des articles publiés (et en pratique malheureusement surtout à son prix). Or cet « impact factor » est très important pour les scientifiques car il est un facteur de recrutement et de reconnaissance. Il est donc nécessaire d'apporter un équivalent à ce chiffre référence.

Compte-tenu des informations disponibles, il sera possible d'inclure dans cette fonction de score :

- La moyenne des notes des relecteurs et des lecteurs.
- Le nombre de citations par d'autres articles.

- Le nombre de téléchargements de l'article.
- La date de publication de l'article.

Cette fonction reste à définir et nécessitera de nombreux tests utilisateurs. Il serait intéressant de tester aussi les différences de notation lors d'un anonymat des articles de façon à garantir une pertinence optimale de la méthode.

3.2.4 Fonction de recherche

L'utilisation de serveurs distants rend le temps de recherche proportionnel au nombre de serveurs et à l'état du réseau. Il est donc inconcevable, pour un tel projet, de parcourir l'ensemble des articles à chaque requête.

Une solution envisageable pourrait être la sous-traitance par un moteur de recherche comme Google³ ou Quaero⁴ qui adapterait ses critères de recherche à la spécificité de la publication scientifique.

Des solutions de recherche plus pointues sont évoquées dans la partie 4.1 au sujet des Web Services et de l'auto-financement.

3.3 Point de vue financier

Il est très difficile d'estimer les besoins nécessaires à la mise en ligne d'un tel services, tout dépend de sa popularité. Le nombre de serveurs devra être déterminé a posteriori, lorsqu'une première estimation de la charge pourra être réalisée. L'hébergement du service pourra être sous-traité et le budget devra alors être recalculé en conséquence.

Le développement de l'application de base est planifié sur 24 mois. Il nécessite 4 développeurs, 1 développeur web/ergonome et 2 administrateurs réseau pour gérer les serveurs. Parallèlement, 1 personne devra être chargée de la communication autour du service. Cette personne devant compter de nombreux contacts, cette tâche peut être assignée au directeur de projet.

Le service une fois lancé devra être maintenu et mis à jour régulièrement pour proposer les diverses fonctionnalités évoquées ensuite. Ce travail pourra être effectué par la même équipe. Le budget annuel est donc pour les 8 salariés d'environ 50 000 euros⁵ auxquels il faut ajouter le prix des serveurs et des machines de développement soit environ 25 000 euros⁶ si l'on se base sur 8 machines et 4 serveurs pour commencer. Cela représente donc 75 000 euros par an⁷, soit 375 000 euros si l'on se base sur un projet de recherche européen de 5 ans.

³<http://google.com> participant déjà à <http://scholar.google.com>.

⁴Futur moteur de recherche européen, plus d'informations sur Wikipédia : <http://fr.wikipedia.org/wiki/Quaero>.

⁵Sur la base de 5 000 euros par mois et par personnes, toutes charges comprises.

⁶Soit environ 2 000 euros par machine.

⁷On considère ici le renouvellement des machines tous les deux ans, le budget leur étant attribué les autres années servant à financer les voyages, conférences, ... du chargé de communication.

4 Perspectives

4.1 Auto-financement

L'un des premiers objectifs à atteindre, outre le fait de fournir un service de qualité, est d'essayer de l'auto-financer. Plusieurs options peuvent à ce titre être envisagées :

- Faire une campagne de dons.
- Proposer des Web Services spécifiques payants.
- Demander une participation internationale.
- Demander une participation de nature matérielle aux grands groupes susceptibles de concéder des serveurs (IBM, HP, ...).
- Établir un système de partenariat avec d'autres sites.

Les besoins financiers n'étant pas astronomiques, il est probable que de telles mesures permettent un auto-financement du service. Ce site ayant pour but d'être non commercial, il faudra décider de la politique à adopter si le rapport $\frac{\text{dons}}{\text{dépenses}}$ devient excédentaire.

4.2 Ouverture du code

Après lancement, il pourrait être intéressant de placer le code de l'application développée sous licence libre et de proposer aux personnes motivées de participer bénévolement à son évolution. Dans ce but, il sera nécessaire de placer le code source sur un Système de contrôle de Versions Concurrentes (CVS) comme Subversion¹ ainsi qu'un système de triage de bugs comme Bugzilla². Enfin pour favoriser la communication au sein de l'équipe de développement et vis-à-vis de son interaction avec des développeurs « extérieurs », une liste de diffusion sera mise en place. Cette méthode fiable est celle adoptée actuellement pour le développement de Logiciels Libres à l'échelle mondiale.

4.3 Extension des services liés aux comptes

L'ajout de fonctionnalités peut sembler futile, voire nuire à la clarté du service si cet ajout n'est pas réalisé de manière non intrusive. Pourtant, certains des services ci-dessous sont très utiles dans un travail de recherche collaboratif entre plusieurs équipes par exemple, que ce soit au niveau du

¹<http://subversion.tigris.org/> et son interface courante Trac : <http://www.edgewall.com/trac/>.

²<http://www.bugzilla.org>.

partage des connaissances selon un certain degré de confidentialité ou lors de la rédaction d'articles.

Plusieurs services actuellement en ligne ont déjà atteint un niveau de popularité satisfaisant et seraient intéressants dans le cadre de ce projet de recherche :

- **Writely**³ permet de rédiger des documents collaboratifs en ligne tout en ayant les fonctionnalités d'un traitement de texte usuel.
- **Voo2do**⁴ permet de tenir des *listes de choses à faire* (todo-list) personnelles et collaboratives permettant par exemple dans le cadre de la recherche d'attribuer les différentes tâches à son équipe.
- **CalendarHub**⁵ permet de tenir un calendrier personnel ou collaboratif, les différents niveaux d'accès permettront de créer un agenda des événements scientifiques par région ou par discipline par exemple.
- **Doodle**⁶ permet de fixer des dates de rendez-vous lorsqu'il faut rassembler plusieurs personnes dans le cadre d'une réunion ou d'une conférence par exemple.
- Enfin l'objectif étant de rassembler l'ensemble de ces services sur une même plateforme, il serait intéressant de les rendre interactifs, à la manière de **Basecamp**⁷ qui permet une gestion de projet collaborative avec l'implémentation d'agendas, de *listes de choses à faire*, d'assignation des tâches,...

Il n'est pas exclu que cette plateforme héberge des débats thématiques ou des visio-conférences dans un second temps.

4.4 Internationalisation

L'un des objectifs à long terme est l'internationalisation de l'interface du site et parallèlement la capacité d'intégrer des articles scientifiques dans des langues différentes de l'anglais. Il faut donc développer ce service en conséquence pour permettre aux visiteurs de traduire l'interface du site dans leur langue maternelle de façon aisée, Rosetta⁸ pourrait être utilisée en ce sens.

³<http://www.writely.com>.

⁴<http://voo2do.com>.

⁵<http://calendarhub.com>.

⁶<http://www.doodle.ch>.

⁷<http://basecamp.com>.

⁸<http://launchpad.net/rosetta>.

5 Conclusion

Comme l'énonçait J. H. Poincaré il y a plus d'un siècle :

La liberté est pour la Science ce que l'air est pour l'animal¹.

J'ai choisi ce sujet car je pense que la Science doit consacrer ses fonds à l'accroissement du savoir collectif. Or, le mécanisme actuel de publication ne permet pas d'avoir accès à ce savoir de façon libre et constitue, pour les chercheurs, une importante perte d'argent et de temps, ce dont ils manquent cruellement.

L'implémentation du service décrit permettrait de bénéficier, en plus d'une interface commune de dépôt et d'accès aux articles, d'une véritable plateforme scientifique de travail améliorant les conditions de collaboration entre les différentes équipes scientifiques.

Néanmoins, ce projet, par son ampleur, sera difficile à mettre en œuvre. Il ne nécessite pas seulement la motivation d'une équipe mais les choix politiques de nombreux pays. De plus, il est basé sur la participation active des chercheurs, aussi bien au niveau du dépôt des articles que de la notation et de la relecture. C'est un risque mais au vu de la popularité grandissante des wikis sur internet², la participativité de l'internaute fera bientôt partie intégrante de sa navigation sur la toile ce qui me permet d'être optimiste à ce sujet.

Un autre problème qui n'a pas été abordé dans ce rapport est l'enjeu économique que représente l'industrie de la publication scientifique. Il a été estimé à 18.4 milliards de livres sterling par an (soit 27 milliards d'euros) et à l'origine de 164 000 emplois pour le Royaume Uni uniquement³. On comprend mieux devant de tels chiffres les pressions qui doivent peser sur le gouvernement dans le choix d'une publication scientifique libre. . .

Hormis ces contraintes d'ordre politique, ce projet est techniquement tout à fait réalisable et constituerait une réelle avancée vers l'accès des résultats de recherche pour tous.

¹Jules Henri Poincaré (1854-1912), mathématicien et physicien français.

²Dont l'exemple le plus connu est wikipédia : <http://fr.wikipedia.org>.

³<http://www.publications.parliament.uk/pa/cm200304/cmselect/cmsctech/399/39905.htm>.