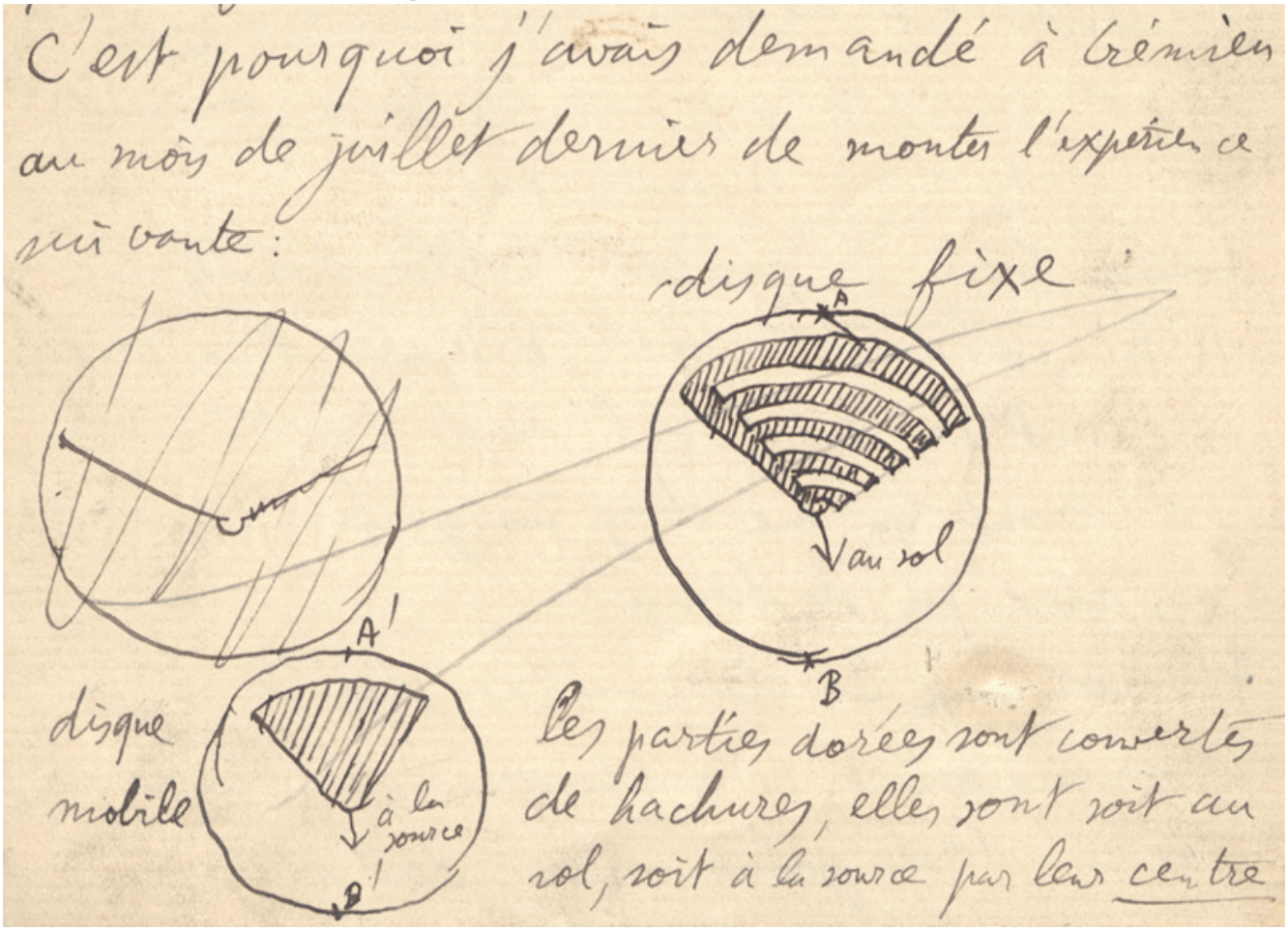
Partage d'expériences

L'histoire des sciences pour les robots : les humanités numériques aux Archives Henri Poincaré



Détails d'une Lettre de Poincaré à A. Potier, janvier 1901. Une transcription de cette lettre se trouve à l'adresse suivante : http://www.univ-nancy2.fr/poincare/chp/text/potier20.xml © 2014 Henri-Poincaré Archives

Depuis l'apparition des moteurs de recherche, les chercheurs ont trouvé de nouveaux lecteurs, pas comme les autres. Ces nouveaux lecteurs sont des robots, des automates qui parcourent les arborescences des sites web. Pour ceux qui cherchent à accroître leur lectorat, ces robots deviennent des lecteurs privilégiés. En effet, les moteurs de recherche dirigent de plus en plus les lecteurs humains vers les documents électroniques indexés lors des recherches systématiques de l'espace en ligne par les robots. Or, un problème de fond se pose ici, à propos de l'écriture. À l'école, on apprend bien comment écrire pour communiquer avec les lecteurs humains mais, pour l'instant, on n'y apprend pas comment écrire pour les robots, à moins d'être roboticien.

Les choses seraient plus simples, naturellement, si les robots comprenaient le langage naturel des humains. Ils font déjà des prouesses d'indexation de données textuelles, à l'aide des linguistes, notamment. Cependant, dans l'attente d'une indexation convenable, automatique et passive des documents en ligne, nous devons envisager comment informer les robots du contenu de nos documents. On sait que les appareils photo numériques enregistrent des métadonnées en même temps que l'image numérique, de telle sorte que les robots puissent indexer chaque fichier correctement. D'une manière semblable, ce qu'il nous faut est un outil qui nous permette de créer nos textes sans nous soucier de leur lecture par les robots, avec enregistrement automatique de métadonnées.

Dans le cadre de ce « partage d'expérience », j'aimerais présenter des exemples concrets de cette problématique des humanités numériques dans le contexte d'un laboratoire d'histoire des sciences et de philosophie: les Archives Henri Poincaré (LHSP-AHP, UMR 7117, CNRS / Université de Lorraine). Je m'intéresse à deux types de projet numérique en particulier : la gestion et l'exploitation de fonds en histoire et philosophie des sciences et le projet d'annotation critique des papiers d'Henri Poincaré. En fin d'article, je livre quelques chiffres de fréquentation de nos sites web.
Avant d'aborder ces sujets, il convient de rappeler brièvement l'évolution des fonds concrets et virtuels aux Archives Poincaré

depuis la fondation du laboratoire en 2000. Le laboratoire possédait alors une reproduction microfilm des archives de la famille Poincaré, ainsi que deux milles tirés à part, concernant surtout la logique mathématique après 1930. Par la suite, le laboratoire a acquis plusieurs fonds personnels, dont ceux des philosophes Louis Couturat, Louis Rougier, Jules Vuillemin, du psychologue Alfred Binet, du mathématicien Pierre Dugac et du groupe Bourbaki. Les archives de la famille Poincaré ont été numérisées en 600 dpi, avec des centaines de manuscrits de Poincaré déposés dans les archives privées et publiques en Europe et aux États-Unis. À ces fonds personnels s'ajoutent deux fonds virtuels et trois catalogues bibliographiques, constitués dans le cadre de projets de recherche variés. Les sites publiés par les chercheurs du laboratoire sont tous accessibles en ligne.

Gestion, mise en valeur et partage du patrimoine

La nature hétéroclite des fonds aux Archives Poincaré, aussi bien par rapport aux objets que par rapport à l'exploitation qu'on en fait actuellement, ou que l'on prévoit d'en faire dans les années à venir, implique une approche inclusive. Le recrutement d'un ingénieur d'études, Pierre Couchet, spécialisé dans la publication électronique, nous a permis de choisir une solution adaptée à nos besoins : la plate-forme Omeka, développée par le *Rosenzweig Center* à *George Mason University*.

Un grand avantage de la plate-forme Omeka, au moins pour ceux qui voudraient privilégier la communication avec les robots, est l'exposition automatique des métadonnées. A cela s'ajoute la simplicité d'utilisation et la construction LAMP (Linux, Apache, MySQL, PHP) d'Omeka ainsi que l'existence d'une communauté importante d'utilisateurs[1].

Nous avons également pris en considération la question de l'intégration des fonds existants. Parmi les deux fonds numérisés aux Archives Poincaré – les corpora Poincaré et Bourbaki – seul ce dernier a été porté sur Omeka. La raison principale en est que le fonds Bourbaki est composé surtout d'images et de métadonnées et se prête ainsi à l'intégration dans Omeka, alors que le corpus Poincaré comprend en outre des milliers de transcriptions en LaTeX, avec un système de gestion *ad hoc*.
Avec les sites Poincaré et Bourbaki, deux bases de données ont été publiées sur le serveur de l'Université de Lorraine avant l'adoption d'Omeka : la Bibliographie de Poincaré et les Nouvelles Annales de Mathématiques. Le développement par Pierre Couchet de ces applications en LAMP nous a montré les avantages de ce type de construction. Les responsables de ces sites apprécient surtout la possibilité de mettre à jour en toute sécurité les données de ces bases disponibles au grand public.

Une troisième base de données, le Répertoire bibliographique des sciences mathématiques (1894-1912), a été élaborée par Laurent Rollet en collaboration avec la Cellule MathDoc. Ce répertoire ancien, construit par des mathématiciens à la fin du XIX^e siècle, figure parmi les bases de données bibliographiques sur le site de MathDoc ; il est hébergé par la Cellule MathDoc.

Au-delà de l'intégration des fonds anciens, c'est surtout dans la facilité de construction de nouveaux sites web que l'on peut voir un avantage de la plate-forme Omeka. Aux Archives Poincaré, de nombreux sites et bases de données figurent sur la plate-forme Omeka "ahp-numérique", hébergée par la TGIR Huma-Num. Parmi ces sites, deux concernent les institutions scientifiques de Nancy : HISE et Archives de la Faculté des sciences de Nancy (AFSN).

Alors qu'ils traitent de la même histoire, les deux sites HISE et AFSN se distinguent par leur approche documentaire. Le site AFSN est purement textuel ; à travers le résumé d'arrêtés déposés aux archives de l'université, il permet de suivre la gestion des ressources humaines et les carrières menées à la Faculté des sciences de Nancy, des doyens, directeurs d'institut, enseignants, agents comptables et personnel administratif et technique, des années 1870 aux années 1960. Le site HISE, en revanche, donne accès à des numérisations de documents divers, dont des procès-verbaux de la Faculté des sciences de Nancy et des photographies d'instruments. Un titre et des mots-clés sont associés à chaque image, ce qui facilite leur localisation, aussi bien par les lecteurs humains que par les robots (à travers l'exposition par Omeka des métadonnées).

Annotation critique du corpus Poincaré

La question de la lecture automatique se pose différemment selon qu'il s'agit de l'édition électronique de la correspondance, des images ou des documents divers avec ou sans formules mathématiques. Le contenu des lettres échangées entre Poincaré et ses pairs est indexé par les robots, ce qui fait qu'une suite de mots-clés envoyée à un moteur de recherche suffit souvent pour localiser une lettre parmi les centaines de millions de documents disponibles sur Internet. Pourtant, les moteurs de recherche sont inutiles lorsqu'on cherche, par exemple, une formule mathématique précise, comme on en trouve par milliers dans les papiers de Poincaré. De même, si on cherche une illustration ou un diagramme précis dans ce corpus, les moteurs de recherche ne trouveront rien, à moins que l'objet en question ne soit annoté, ses métadonnées communiquées à un dépôt adéquat, et moissonnées par les robots.

En ce qui concerne les formules mathématiques en particulier, leur publication en ligne est facilitée par l'utilisation de MathML (*Mathematical Markup Language*). Alors que MathML gagne la confiance des auteurs depuis sept ans, il reste deux obstacles à sa généralisation. Pour commencer, les navigateurs n'affichent pas tous correctement les formules codées en MathML. Ensuite, écrire directement en MathML est extrêmement fastidieux.

Ces obstacles peuvent être surmontés, d'abord en précisant aux lecteurs quels navigateurs affichent correctement MathML (comme Google Chrome). Ensuite, depuis fin 2013, on peut rédiger ses documents en LaTeX et les convertir en XHTML avec LaTeXML, un logiciel rédigé en langage Perl. Développé par l'institut américain NIST dans le cadre de la DLMF (*Digital Library of Mathematical Functions*), ce nouveau logiciel libre offre en outre la possibilité de générer des métadonnées RDFa (*Resource Description Framework in attributes*). Ainsi, à l'aide de LaTeX–LaTeXML, on peut produire des métadonnées facilement en RDFa non seulement pour les formules mathématiques de son document, mais aussi pour tout élément XML.

En principe, avec LaTeXML, la sémantique du *markup* LaTeX pourrait être encodée automatiquement dans les métadonnées RDFa. Cette fonctionnalité est prévue par les développeurs de LaTeXML qui soulignent, en passant, qu'il s'agit d'un *work in progress*. Le code source de LaTeXML étant disponible sur GitHub, on peut développer soi-même les fonctionnalités souhaitées. Avec le tandem LaTeX–LaTeXML, on se rapproche de l'exemple de l'appareil photo

1. Pour une présentation des qualités d'Omeka, voir la présentation de Pierre Couchet.

qui enregistre une série de métadonnées dans un fichier image, sans intervention ou effort de la part du photographe.

En tant qu'exemple de l'annotation "automatique" effectuée par LaTeXML, prenons une formule mathématique, comme celle que Poincaré a écrite au mois de janvier 1901 (§48.12), dans une lettre envoyée à son collègue et ancien professeur de physique Alfred Potier (1840-1905). Dans le cadre de son analyse des expériences d'un élève de Gabriel Lippmann, Victor Crémieu, qui cherchait à mettre en évidence l'effet magnétique d'une charge électrique en mouvement (ou l'effet de Rowland), Poincaré a fait appel au théorème de Stokes, afin d'arriver à l'expression

$$\int (\ell A + mB + nC) d\Omega.$$

Pour comprendre la démarche de Poincaré, le lecteur de la Correspondance de Poincaré veut savoir si Poincaré a jamais employé une analyse semblable, par exemple, lors de ses échanges avec Potier et d'autres physiciens ou peut-être à l'occasion d'un cours de physique mathématique ou dans les pages d'un journal. Grâce à la représentation "cachée" en LaTeX, produite automatiquement par LaTeXML à partir de la source en LaTeX et exposée aux robots, ce type de question peut trouver une réponse. Il suffirait de chercher la représentation en LaTeX de la formule en question :

```
\int (\ell A + mB + nC) d\Omega
```

en prenant en compte, éventuellement, les variantes alphabétiques. Ainsi, la représentation en LaTeX qui est employée afin d'afficher la formule pour le lecteur humain sert à distinguer l'objet mathématique visé et, par conséquent, à le rendre accessible à l'indexation.

En somme, l'emploi du couple LaTeX–LaTeXML pour la transcription de la correspondance de Poincaré, satisfait à trois critères essentiels de l'édition critique menée par les chercheurs aux Archives Poincaré. D'abord, le résultat est conforme aux recommandations de la TEI (*Text Encoding Initiative*). Ensuite, la présentation des formules mathématiques est impeccable au niveau de la lisibilité sur écran. Enfin, il permet à l'équipe éditoriale de se concentrer sur la compréhension et l'annotation des textes du corpus Poincaré et de rédiger ses textes en vue des lecteurs humains, tout en réalisant des documents conformes aux standards de l'édition critique et lisibles par les robots.

Chiffres de fréquentation

Afin de cerner l'effet de l'exposition de métadonnées sur le nombre brut de requêtes de page, il faudrait disposer des données quantitatives de ces requêtes avant et après une telle exposition. Or, ces données nous manquent, hélas. Mais à leur place, je propose les chiffres de fréquentation des deux sites mentionnés en amont : ahp-numérique (depuis mai 2013, Figure 1) et la Correspondance de Poincaré (depuis mai 2013, Figure 2). Sont comptabilisées pour celui-ci uniquement les requêtes de pages de transcription en XML et PDF, à l'exclusion des numérisations de manuscrits. On voit que ces deux sites attirent chaque mois des milliers de requêtes de page, jusqu'à 24 000 requêtes en juillet 2013, pour la correspondance de Poincaré. Si on comptait les robots, ce chiffre serait plus élevé, sans doute.

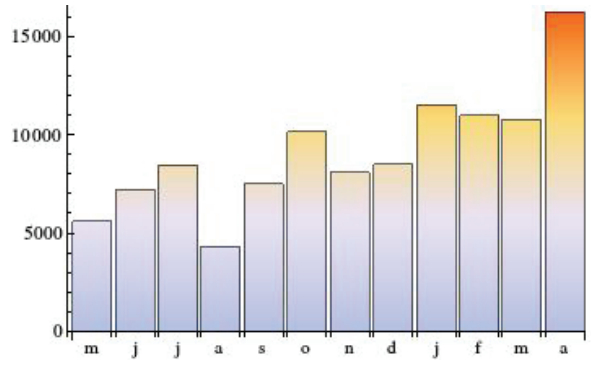


Figure 1. Requêtes de page du site ahp-numérique depuis mai 2013.

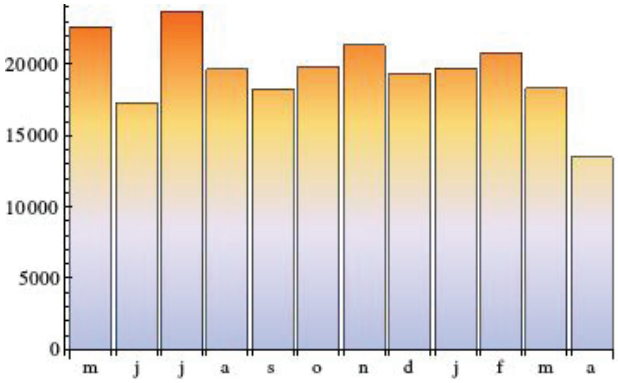


Figure 2. Requêtes de page du site de la Correspondance de Poincaré depuis mai 2013.

contact&info
▶ Scott Walter
scott.walter@univ-lorraine.fr
▶ Pour en savoir plus
http://henri-poincare.ahp-numerique.fr/

contact&info
▶ Nadine Dardenne
Chargée de la communication
et de la structuration des réseaux
nadine.dardenne@huma-num.fr
▶ Pour en savoir plus
www.huma-num.fr