Numérisation des publications des IREM

Jean-Louis Maltret

14 juin 2011

1 Contexte

Les dernières années ont été marquées par le fait que les ressources utilisées par les enseignants sont de plus en plus souvent des ressources numériques en ligne, qu'elles soient sur des sites institutionnels, associatifs ou personnels.

Le réseau des IREM, producteur de resources depuis 30 ans, accuse de ce point de vue un certain retard. Les productions nouvelles sont majoritairement en ligne, mais les productions anciennes, qui sont encore souvent d'actualité, sont en format papier, qui plus est dans certains cas absolument introuvables.

Il semble donc indispensable de lancer une opération de numérisation et mise à disposition des ces ressources. L'objet de cette note est d'en préciser les différents aspects et les modalités éventuelles.

2 Déroulement

Les publications récentes ont en général une archive numérique, ce qui est à vérifier et à impulser si ce n'est pas fait de façon correcte. Il n'est donc nécessaire de numériser qu'en deça d'une date qui est environ 2000, parfois plus tard si l'archivage n'a pas été fait de façon satisfaisante. Compte-tenu des informations contenues dans les catalogues réalisés à l'Irem de Lyon (publications à partir de 1971) on peut estimer à 5000 le nombre de brochures publiées. Ceci est cohérent avec les fiches de `publimath` sur les publications Irem : il y en a 2591 correspondant à la période 1972-2000, et la proportion de 1/2 pour les fiches réalisées est plausible. Si on prend 10 pages comme taille moyenne d'une brochure on a une estimation de 50.000 pages à numériser.

Les étapes du processus peuvent se détailler comme suit :

1. le passage au scanner, création d'un fichier image (pdf ou autre format)
2. la reconnaissance de caractères pour avoir un document susceptible de recherches (format pdf-a ou djvu, analogue à ce qu'on a sur Numdam)
3. les corrections d'erreurs et la mise en forme éventuelle
4. l'intégration dans une base de données
5. le catalogage type `publimath`
6. la présentation et l'interfacage pour utilisation publique

Les étapes 1-2, ont été présentées dans la vidéo réalisée par Olivier Roizés, disponible à http ://vimeo.com/22618385 mot de passe : Publimath. Elles peuvent être réalisées de plusieurs façons (sous-traitance comme Numdam, réalisation intégrée, étapes découplées,...) mais représentent dans tous les cas un coût en personnel ou en budget (ordre de grandeur un peu moins d'un euro par page si on le sous-traite).

L'étape 3, beaucoup plus fastidieuse, conditionne la qualité du résultat final : une erreur de reconnaissance entre "o" et "a" aboutira ultérieurement à une indexation fausse et à des recherches infructueuses, sans parler du traitement des formules, indispensables pour les textes à traiter. Les publications anciennes ne sont pas toujours de bonne qualité typographique et engendreront certainement des erreurs de reconnaissance. Cette phase n'est pas automatisable et a un coût humain important. Les entreprises qui le font ont des coûts prohibitifs et n'auront sûrement pas la compétence mathématique requise.

Les étapes 4-5 sont en partie réalisées par `publimath` et entrent tout à fait dans le développement des travaux de la commission.

L'étape 6 nécessite des choix techniques (données centralisées ou pas) et serait donc l'occasion de relancer le processus qui avait été engagé avec publirem/navirem.

3 Moyens

Il faut distinguer ce qui est centralisé et ce qui est réparti. Si, pour être efficaces, l'indexation et la recherche doivent être centralisées, le stockage et la sauvegarde des fichiers pourraient être répartis dans les Irem (pas nécessairement tous vu l'hétéroénéité des moyens disponibles). Dans tous les cas le volume de données (milliers de fichiers, beaucoup plus si des brochures sont découpées en articles) impose une rigueur de gestion qui représente un travail conséquent dont la commission `publimath` a l'expérience.