

LA BIBLIOTECA MATEMÀTICA DIGITAL

JAUME AMORÓS

La Biblioteca Matemàtica Digital és un somni, que ha esdevingut un projecte debut als avanços de la tecnologia informàtica dels darrers anys.

El somni és universal i clàssic:

- que cadascú tingui accés a tota la literatura matemàtica de tots els temps,
- que aquest accés sigui ràpid en la localització i obtenció dels documents, amb la possibilitat de fer cerques segons el tema dels articles o llibres,
- que el cost de tot plegat sigui assequible.

D'aquest somni en diríem també la Biblioteca Ideal, i les biblioteques universitàries ja s'hi han acostat bastant: grans col·leccions de revistes i llibres matemàtics, catàlegs públics, serveis de fotocòpia i préstec interuniversitari que permeten obtenir qualsevol article de recerca del món, a més de nombrosos llibres. La digitalització de les bases de recensions Mathscinet i Zentralblatt fa possible des de fa pocs anys fer cerques d'articles segons el seu contingut, tot i que incompletes.

Qué ens falta en aquest punt per arribar a la biblioteca ideal?

- Els usuaris del sistema de biblioteques tenim dret a fotocopiar els articles però no els llibres, fins i tot si estan exhaurits.
- El creixement exponencial de la literatura matemàtica implica un increment també exponencial dels costos d'adquisició de nous textos i manteniment de la col·lecció.
- Si la nostra institució no té el recurs que volem, tardarem molt en aconseguir-lo.

La definició inicial d'aquest article era un farol. El projecte de la Biblioteca Matemàtica Digital no serà la biblioteca ideal que voldríem, sinó una aproximació més bona que l'existent. Es basa en treure partit de les noves tecnologies digitals, pensades com un mitjà en l'avanç cap a la biblioteca dels nostres somnis. L'assoliment de la situació ideal descrita més amunt no dependrà només de que introduïm més tecnologia i organització en el nostre arxivatge: caldrà també que els

Date: 15 de Juny del 2003.

Aquest article es deu a tasques encomanades per la Societat Catalana de Matemàtiques a l'autor. Lluís Anglada, del CBUC, i Quique Macías, de la U. de Santiago, han proporcionat informació valuosa.

matemàtics tinguem en compte el cost econòmic i les restriccions legals a l'hora de difondre les nostres obres.

1. EL PROJECTE DE LA BIBLIOTECA MATEMÀTICA DIGITAL (DML)

Aquest projecte neix de converses entre P. Tondeur i J. Ewing, aleshores director de la DMS, NSF i president de l'AMS respectivament, l'any 2001 (vid [5]).

L'objectiu inicial és coordinar els projectes de digitalització de literatura matemàtica que ja aleshores començaven a funcionar, difonent *estàndars tècnics i bones pràctiques* per a fer que aquests projectes siguin compatibles entre si i marxïn en la direcció de la biblioteca ideal.

Un objectiu complementari és promoure projectes de digitalització allà on no n'hi havia. Donada l'enormitat de la tasca ([5] estima 50 milions de planes per digitalitzar), hom proposa una organització de projectes *en arbre*: projectes nacionals, editorials i biblioteques són els que porten a cap la digitalització de la literatura. El projecte mundial i les seves branques principals, de moment la nord-americana i l'europea, recolzaran aquests projectes i en difondran els resultats. Hi haurà un catàleg a Gotinga, que si tot va bé acabarà sent el catàleg universal de les Matemàtiques.

La branca catalana del projecte de Biblioteca Matemàtica Digital neix a l'hivern 2002–3, simultàneament amb la branca europea (projecte DML-EU, que es sol·licita a la Unió Europea dins del 6^e Programa Marc aquesta primavera). A fi d'aconseguir una categoria de soci per sobre de la nostra experiència digitalitzadora, participem en el projecte DML-EU dins d'un consorci amb la branca espanyola del projecte, que impulsa la RSME. Les sis revistes matemàtiques catalanes indexades internacionalment han acceptat participar en el projecte; en els primers anys aquest consistirà en digitalitzar els fons en paper d'aquestes revistes i de les seves predecessores, posant-los a la web amb accés universal. Un cop assolit aquest objectiu, hom preten que el projecte continui cap *endavant*, amb les revistes afegint els seus articles al repositori a mesura que cedeixin al públic el dret de còpia. S'estudiarà la possibilitat de continuar-lo cap endarrera, si hom disposa de fons bibliogràfics susceptibles de participar i de finançament.

2. QUÉ ÉS LA DIGITALITZACIÓ?

Donem una descripció sucinta dels aspectes tècnics d'aquesta empresa.

Pels textos que hom té només en paper, anomenarem al pas inicial *primera digitalització*. Consisteix en l'escaneig de l'original en paper complint uns requisits mínims de qualitat, bastant baixos debut a la rellevància dels gràfics i qualitat d'impressió habituals, i uns altres mínims d'arxivatge, més exigents, pensats per a que en el futur hom

disposi de les *imatges digitals originals* per fer versions dels documents en formats més avançats/alternatius.

Denotarem *segona digitalització* la fase que segueix: a partir de les imatges obtingudes a la primera digitalització, hom prepara un fitxer multipàgina que contingui l'obra en un format convenient per veure per pantalla/imprimir (avui pdf, demà un altre). Aquest fitxer conté també *metadades*: camps de text consultables de manera personal o automàtica, amb informació bibliogràfica i sobre el contingut de l'obra.

Aquesta segona fase és a priori la més difícil, però els matemàtics disposem d'un recurs que és un autèntic tresor: les bases de recensions Mathscinet (de l'AMS) i Zentralblatt (de Springer-Verlag). Aquestes dues bases de dades estan completament digitalitzades: si hom coneix algunes paraules clau bibliogràfiques d'una obra digitalitzada (per haver fet un reconeixement parcial del text, o per haver-les introduït a ma), podem identificar-la en alguna d'aquestes bases i obtenir així no només la informació bibliogràfica que ens faltava, sinó també una recensió sistematitzada del contingut! Hem d'agraïr tant a l'AMS com a Springer-Verlag la seva col·laboració plena en aquest projecte.

La recensió d'un article pot ser des de millor que l'obra original fins poc informativa. El projecte europeu DML-EU adreça aquesta qüestió mitjançant les eines informàtiques de reconeixement de textos (OCR): hom proposa, usant l'experiència en reconeixement de textos, multilingüisme i traducció tècnica informatitzada que ha estimulat la unificació europea, fer un reconeixement del text dels articles, i adjuntar-lo com un camp de text al fitxer preparat en la segona digitalització. No es pot fer un programa pdf2tex perquè és impossible reconèixer les fòrmules, i fins i tot els millors programes d'OCR tenen una taxa d'error de l'ordre del 0.5%. Però un camp de text que contingui tot el text de l'obra, amb taxa d'error entorn a l'1 % en les paraules d'un diccionari, i la possibilitat d'assenyalar on apareixen les paraules en el fitxer digital, permetrien una millora dramàtica en la cerca de resultats específics en la literatura.

3. ASPECTE LEGAL: EL DRET DE CÒPIA

El dret de còpia de les obres matemàtiques és una qüestió de la màxima importància per organitzar una Biblioteca Ideal.

En el cas dels articles i de la majoria de llibres moderns, aquest dret de còpia pertany als editors que han publicat les revistes o els llibres. Hi han limitacions a aquesta propietat: de temps, si l'autor és realment antic, i, sobre tot, la possibilitat de fotocopiar articles de les revistes a la biblioteca. L'ordenació legal d'aquesta pràctica sempre ha estat poc clara pels usuaris pel que fa a consentiments i pagament de drets, però està indubtablement tolerada pel que fa a còpies per ús propi, i arriba lluny via els sistemes de consulta o préstec interbibliotecari.

Les tecnologies digitals són una revolució en curs en el camp de l'edició científica: fins i tot si hem de fer la primera còpia digital en un escanner/fotocopiadora, totes les còpies digitals successives tenen cost econòmic i de qualitat zero.

Aquesta evolució provoca un debat legal de primer ordre: la Unió Europea té una nova directiva sobre el dret de còpia en la societat de la informació, que s'està traslladant a les legislacions nacionals dels membres. El projecte de llei que es debat a Alemanya autoritza a les biblioteques públiques a fer còpia digital dels seus fons en paper, i a posar-los a disposició dels seus usuaris (tothom amb accés a Internet). Les editorials, protesten vigorosament i es queixen de que una disposició així les pot arruïnar.

La política del projecte de Biblioteca Matemàtica Digital en aquest tema es basa en el respecte escrupulós del dret de còpia dels editors, ja que una component essencial d'una Biblioteca Ideal és que creixi de manera ordenada amb el nou material que es publiqui, i les revistes són avui en dia les publicadores de la recerca.

L'acció que s'empren en aquest tema és intentar organitzar l'interés dels matemàtics, com a productors i alhora consumidors de la nostra literatura, per a buscar un compromís entre la viabilitat del negoci editorial i l'assequibilitat de les obres. Pel que fa a les revistes, per a ser comptades com part de la Biblioteca Matemàtica Digital han de seguir un sistema batejat *moving wall*: els articles són disponibles només pels subscriptors els x anys posteriors a la seva publicació, i després són universalment accessibles. Els terminis de *moving wall* acceptats per les revistes incorporades al projecte solen ser de fins a 5 anys quan l'editor és una universitat o societat científica (com ara l'AMS), o de fins 10 anys en el cas d'editorials privades (per exemple Springer-Verlag). Les revistes matemàtiques catalanes han acceptat adoptar sistemes *moving wall* per l'accés als seus articles, i estan perfilant els seus terminis.

4. MATEMÀTIQUES DE FRANC A LA WEB

Editorials, universitats i biblioteques van començar cap a l'any 1997 a posar els seus fons digitals a la web. Aquests esforços no estan tan coordinats com la Internet permetria (la web de la revista encara no apunta a col·leccions seves antigues ...). Podem classificar-los segons un pla afí $(\mathbb{Z}/2\mathbb{Z})^2$: en un eix tenim publicacions noves, *nascudes digitals*, vs fons antics escanejats, en l'altre tenim accés universal vs accés només per subscriptors.

Assenyalem alguns projectes de digitalització amb contingut en xarxa que poden interessar als lectors catalans:

- JSTOR (www.jstor.org): Té unes dotze revistes de Matemàtiques, entre elles *Annals of Mathematics*, *Econometrica*, les revistes de l'AMS i de la SIAM, des de l'inici de cada revista fins el 1997.

Accessible només per subscriptors. A Catalunya UAB,UPC, UPF ho són.

- Euclid (<http://ProjectEuclid.org>): 19 revistes de qualitat reconeguda, s'ofereixen per la web amb una interfície unificada, però amb polítiques d'accés diverses (de l'universal a tot per subscriptors), i amb un nombre d'anys en oferta també divers.
- L'AMS (www.ams.org/journals) ofereix directament les seves revistes a partir del 1997, amb moving wall de 5 anys.
- El projecte NUMDAM (www.numdam.org) forma part de la digitalització del patrimoni cultural francès, i ofereix 4 de les millors revistes franceses, amb moving walls diversos.
- El Centre de Digitalització de la Biblioteca de Gotinga (GDZ, <http://gdz.sub.uni-goettingen.de/en/index.html>) ha digitalitzat nombrosos llibres vells de matemàtiques, més un surtit interessant de revistes modernes, en col·laboració amb Springer-Verlag, i les posa amb accés universal (exemples: *Mathematische Annalen 1869–1996*, *Inventiones Mathematicae 1966–1996*).

Aquesta llista de projectes i revistes augmentarà constantment els propers anys. Per informació d'accés més actualitzada consulteu la web del projecte DML català ([3]) o la del meta-projecte EMANI ([4]).

REFERENCES

- [1] Web del projecte mundial DML: <http://www.library.cornell.edu/dmlib>
- [2] Sol·licitud de projecte europeu DML-EU, declaració d'intencions de l'EMS: <http://www.library.cornell.edu/dmlib/DML-EoI-draft6.pdf>
- [3] Web sobre digitalització matemàtica a la SCM: <http://www.iecat.net/scm> (anar a l'apartat *Biblioteca Matemàtica Digital*).
- [4] El catàleg de digitalitzacions a la web: <http://www.emani.org>
- [5] J. Ewing, *Twenty centuries of mathematics*. http://www.ams.org/ewing/Twenty_centuries.pdf
- [6] E. Macías-Virgós, *Un gran proyecto de colaboración internacional: la Biblioteca Digital de Matemáticas*. A aparèixer al Boletín de la RSME.

NOTA ALS EDITORS: POSAR COM REQUADRE A PART EL QUE SEGUEIX

La digitalització dels pobres

Els projectes aquí descrits digitalitzaran gran part de les revistes matemàtiques més usades, però no totes (falten per exemple les de la London Math. Soc.). A més hi ha una massa enorme de literatura apart de les revistes (llibres vells, notes de seminaris ...) amb interès desigual. Podem permetre'ns aquesta digitalització? La resposta és que, com en tantes coses, el cost de la digitalització depèn de la qualitat exigida de manera almenys quadràtica. Vet aquí una estratègia per digitalitzar documents amb un cost monetari zero:

- (i) Les fotocopiadores modernes són escanners: digitalitzem cada vegada que fotocopiem un article. A sobre, tots els models nous tenen port Ethernet per a poder servir com impressores en xarxa.
- (ii) Els projectes de digitalització com el DML-EU estan creant servidors web que rebin el resultat de la primera digitalització (l'escanejat), i facin la segona. En dos o tres anys el nivell de qualitat d'aquest segon pas serà força bo per obres indexades en Mathscinet o Zentralblatt.
- (iii) Els fotocopiadors dels articles, o les biblioteques, podran enviar el resultat de la fotocopia a aquests servidors, que faran la segona digitalització si és legal i la retornaran al remitent.
- (iv) El remitent original farà un control de qualitat molt lleuger a l'article així digitalitzat, i el posarà en una xarxa P2P (com l'antic Napster, Kazaa ...)

Quins problemes té la digitalització del pobre? La qualitat dels escanejats serà molt variable. Per això es recomanable anar posant aquests fitxers en una xarxa P2P, i que els usuaris vagin afegint noves versions fins que les acceptables estiguin prou difoses. Un altre perill és el que corre el dret de còpia de l'editor en un sistema així, però els programes que el projecte DML-EU desenvolupa inclouen la identificació de revista, això permetrà que el servidor de la segona digitalització operi només amb consentiment de l'editor. Finalment, assenyalem l'avantatge d'aquest sistema: amb un cost proper a zero, es digitalitzarien nombrosos documents en ordre estricte d'interés dels usuaris!

DEPT. MATEMÀTICA APLICADA I, UNIVERSITAT POLITÈCNICA DE CATALUNYA, DIAGONAL 647, 08028 BARCELONA
E-mail address: jaume.amoros@upc.es