À travers le miroir d'une bibliographie

Roberto Di Cosmo
DMI-LIENS (CNRS URA 1347)
Ecole Normale Supérieure
45, Rue d'Ulm
75230 Paris France
Email :dicosmo@ens.fr
WWW :http://www.dmi.ens.fr/∼dicosmo

19 Mai 1997

Forme et contenu. Ces deux notions ont entretenu plein de philosophes au cours des siècles dans des débats fort intéressants. Est-ce la forme la seule réalité, vu qu'on ne peut pas atteindre le contenu sinon à travers d'une forme particulière dans laquelle il est présenté, ou est-ce seulement une illusion, vu qu'il n'y a souvent pas de forme canonique qui soit préférable aux autres ?

Aujourd'hui, les sciences de l'information, l'informatique, peuvent aussi apporter leur modeste contribution à ces débats : quand on doit gérer un système d'information, la distinction entre le contenu en information et la forme dans laquelle l'information est présentée n'est plus juste une question philosophique, mais ça peut entraîner des pertes (ou des gains) qui se chiffrent en millions de dollars. Dans cette branche de l'informatique théorique qui étudie les bases de données, ça fait très longtemps que la distinction entre forme et contenu a été redécouverte, étudiée et maîtrisée, même si ces techniques pourtant élémentaires ne sont malheureusement pas toujours appliquées dans les logiciels à disposition du grand public, et cela en raison de sombres enjeux commerciaux qui entraînent des pratiques détestables sur lesquelles on aura lieu de s'etendre dans une autre occasion.

Je vais essayer ici de présenter brièvement les raisons de l'intérêt en informatique de cette distinction, en considérant quelques exemples simples, pour arriver enfin a décrire comment on peut appliquer les leçons qu'on peut en tirer au cas de la gestion des bases bibliographiques.

Dans tout système informatique qui gère des bases de données, on est confronté à un problème très relié à la distinction entre forme et contenu : comment va-t-on garder l'information qui nous intéresse (le contenu) ? Il s'agit d'une information qui change dans le temps (la liste des employées d'une entreprise, la liste des publications d'un chercheur, le catalogue des produits d'un grand magasin), qui doit être présentée sous différentes formes à différents utilisateurs (une bibliographie en français ne se présente pas du tout de la même façon qu'une bibliographie en anglais ou en italien :

les conventions de style sont très différentes), et que l'on veut maintenir cohérente à tout moment (il serait très embarrassant pour une entreprise si les prix affichés sur son site Web résultaient différents des ceux à disposition d'un de ses télévendeurs, ou si la liste des publications d'un chercheur pour la même année résultait différente dans deux rapports officiels de la même institution).

Dans les premiers systèmes informatiques, on faisait souvent l'erreur de garder plusieurs copies du contenu, une sous chacune des formes demandés par les différents utilisateurs. En conséquence de cette organisation naïve, il devenait vite difficile, voir impossible de maintenir la cohérence de l'information : on n'avait plus *un* contenu et plusieurs formes, mais un mélange anarchique de contenu et formes. Comme ces erreurs ont coûté très cher, littéralement, on a vite découvert qu'il fallait garder une seule copie du contenu de l'information, pour permettre de la mettre à jour en conservant la cohérence, et écrire à part des programmes de mise en forme qui présentent ce contenu à chaque utilisateur selon ses nécessités. C'est là la naissance des bases de données modernes : on garde le contenu sous une forme très spéciale, appelée *structure logique*, qui ne fait aucune assomption ni sur les supports matériels utilisés pour stoquer l'information, ni sur les formes d'affichage de cette information ; et on fournit à coté des moyens pour interroger la base des donnée en affichant les résultats sous la forme la plus convenable à chaque utilisateurs.

Malheureusement, aucune connaissance n'est durablement acquise, d'autant plus dans le monde de l'informatique commerciale, donc on retrouve de temps à l'autre les mêmes erreurs encore aujourd'hui : sans aller chercher très loin, il y a très peu de temps une grande entreprise dont je tairais le nom à fait un effort formidable pour mettre tout le catalogue de ses produits sur le Web. Mais en lieu de le faire en suivant les règles de l'art, ce qui signifiait écrire un programme qui produit des pages Web automatiquement à partir de la base de données de l'entreprise , ils ont tout fait à la main en quelques mois ; une fois le travail fini, ils se sont aperçu qu'il était impossible de le maintenir à jour automatiquement (la seule façon de le faire à un coût raisonnable). Je suis sûr que si vous utilisez régulièrement le Web vous aurez bien d'exemples d'erreurs semblables qui vous reviennent.

Notons au passage qu'un phénomène analogue, même si pas identique, se présente avec les différents *format* de fichiers utilisés par les programmes d'écriture (word processing) : l'auteur d'un document est souvent pris au piège par ces logiciels encombrants et mal conçus, qui le forcent à mélanger le contenu (ses idées) et la forme (le style de présentation : des mots vont en gras ou en italique, etc.) de façon inextricable. Cela est une source inépuisable de problèmes chaque fois que le même contenu nécessite d'être présenté sous une autre forme, ce qui arrive souvent : un même article peut évoluer au cours des années et paraître dans différentes revues, ou être affiché sur le Web, mais les logiciels d'édition de texte les plus répandus ne permettent pas de sortir leur contenu de leur emprise de façon simple et satisfaisante, et ça nous oblige nous, utilisateur, à perdre un temps fous à pratiquement tout réécrire dans le nouveau format. À nouveau, ce n'est pas ici le lieu de s'intéresser au raisons purement monopolistes et commerciales, et non pas technologiques, de ce phénomène très répandu, mais il est nécessaire de le connaître pour en tirer des leçons qui nous évitent de tomber dans ces

pièges quand on va s'intéresser à la gestion d'une bibliographie.

Venons donc au sujet qui nous est à coeur dans cette petite note : la gestion des informations bibliographique. Essayons d'appliquer les quelques idées que l'on à exposé sommairement ci-dessus : il s'agit de bien séparer contenu et forme dans l'information bibliographique, mais tout d'abord il est important de remarquer qu'il y a *un* contenu et *plusieurs* formes de le présenter, ce qui n'est pas immédiatement évident pour tout le monde. Prenons un exemple : un des mes travaux récents est publié dans l'article suivant

> Di Cosmo (Roberto) et Kesner (Delia). – Combining algebraic rewriting, extensional lambda calculi and fixpoints. *Theoretical Computer Science*, vol. 169, nN^{o} 2, 1996, pp. 201–220.

Ceci devrait paraître tout à fait raisonnable à qui lit ces lignes, vu que c'est la présentation *suivant le style français* de l'information bibliographique. Cependant, ce travail est bien plus souvent cité comme

> Roberto Di Cosmo and Delia Kesner. Combining algebraic rewriting, extensional lambda calculi and fixpoints. *Theoretical Computer Science*, 169(2) :201–220, 1996.

vu que la *lingua franca* de l'informatique reste l'anglais. Quelle de ces deux citations est plus proche du vrai contenu ? Considérez que sur ma page Web vous allez retrouver aussi la version italienne et espagnole, qui sont encore différentes. La réponse, pour un informaticien, est "aucune des deux" : dans les deux cas le contenu est bien présent, mais mélangé de façon inextricable avec des conventions de présentation lié à la langue française ou anglaise (ou italienne, ou espagnole etc.). Extraire le titre, le nom de la revue et les auteurs est possible, mais demande beaucoup d'intelligence et de connaissances sur des conventions arbitraires fixées par chaque langage.

Depuis désormais quinze ans, la plupart des informaticiens décrivent le contenu bibliographique par des enregistrements qui forment une base de données dans laquelle repérer titre, auteur et autres informations est immédiat, et ils choisissent ensuite la forme de présentation (le *style bibliographique*) en fonction de leurs nécessités, laissant à la charge d'un programme automatique, qui est BIBTEX , les détails de la mise en forme.
Dans le cas de ma publication, par exemple, je n'ai écrit à la main ni la version française ni la version anglaise, mais seulement l'enregistrement BIBTEX qui suit :

```
@Article{TCS95,
 AUTHOR = {Di Cosmo, Roberto and Delia Kesner},
 TITLE = {Combining algebraic rewriting,
          extensional lambda calculi and fixpoints},
 JOURNAL = TCS,
 VOLUME  = {169},
 NUMBER  = {2},
```

```
 PAGES   = {201-220},
 YEAR = 1996,
 DMI-CATEGORY = {journal},
 ABSTRACT-URL=
 "http://www.ens.fr/~dicosmo/Publications/Abstracts.html#TCS95.abstract"
}
```

Quelques mots d'explication : la ligne `@ARTICLE{TCS95,` dit que je suis en train de définir une publication que je souhaite nommer `TCS95` pour référence future, et qu'il s'agit d'un `Article` (et non pas par exemple d'un `Book`, qui a besoin d'une mise en page différente). Ce qui suit, c'est le contenu proprement dit : on dit clairement quel est le titre, l'année de publication etc. Même pour l'auteur on n'a préjugé en rien de la présentation : le `and` et la virgule sont là seulement comme séparateurs ; le `and` sépare les auteurs (et il deviendra `et` en français, `e` en italien etc.), tandis que la virgule sert pour identifier de façon univoque le nom et le prénom quand le nom ou le prénom sont composé, comme c'est le cas du nom de celui qui écrit. Un cas intéressant est celui du champ `JOURNAL`, où le mot `TCS` n'est pas le nom complet de la revue, mais une abréviation qui est déclarée ailleurs comme suit

```
@string{TCS = "Theoretical Computer Science"}
```

Cela permet d'imaginer la mise en place d'une base de données de noms de revues qui assure une plus grande cohérence dans les bases de données bibliographiques, et en effet des systèmes semblables sont utilisés depuis des années dans quelques laboratoires d'informatique françaises.

Un atout majeur de ce format est qu'il est *libre*, c'est à dire, le nombre de champs dans un enregistrement n'est pas fixé à priori, mais peut-être étendu à tout moment suivant les nouveaux aspects du contenu qui peuvent se manifester au cours des années : par exemple, le dernier champ `ABSTRACT-URL` est utilisé par un style qui produit des référence à des pages Web, et il n'aurait pas été là il y a quelques années. Tout simplement, si un style bibliographique ne connaît pas ce champs, il l'ignorera.
De même, le champ `DMI-CATEGORY` est utilisé seulement par une classe de styles bibliographiques (décrits dans [1, 2]) qui permettent de produire des bibliographies qui présentent les publications dans un ordre adapté à la rédaction du rapport quadriennal du DMI, mais est ignoré par la plupart des autres styles. Des autres champs que l'on retrouve souvent sont `ABSTRACT` (qui contient un résumé de la publication) et `ANNOTATION` pour des annotations personnelles.

Mais assez de détails : qui est intéressé à utiliser BIBTEX trouvera plein d'informations dans [5, 6, 4] (ces références, bien entendu, ont été générées avec BIBTEX lui-même).

Ce qui compte, est qu'une foi l'information écrite dans ce format explicite, la mise en forme se fera en choisissant le style le plus approprié, qui lui aussi à été programmé *une et une seule fois* (dans l'exemple précédant, l'auteur a utilisé le style `falpha.bst` pour le français et `alpha.bst` pour l'anglais). Il existe aujourd'hui une très large collection de styles dans le domaine publique : avec les prénoms abrégés

ou pas, avec les références Web ou pas, plus un style pour chaque publication majeure en informatique (la revue TCS en utilise un différent de celui de MSCS ou du CACM etc.).

Les avantages de cet approche ne se comptent plus, mais rappellons-en quelques uns

- le contenu est décrit dans un format ouvert et disponible gratuitement
- l'information est écrite une seule fois, et tenue à jour facilement, éventuellement avec l'aide de systèmes de gestion avancée [3]
- si différents auteurs écrivent dans la même revue, et tous utilisent ce format, les bibliographies des différents articles seront parfaitement homogènes et sans demander aucun effort : il suffira de dire à BIBTEX d'utiliser le même format bibliographique pour tous
- il est désormais possible de créer des bases bibliographiques d'utilisation générale : presque toutes les revues ou conférences informatiques ont dressé une liste de publications au format BIBTEX librement disponibles sur le Web, ce qui permet d'avoir des références précises et toujours à jour (une collection de plus de 700.000 entrées est déjà disponible dans `http://wheat.uwaterloo.ca/bibliography/index.html` pour l'informatique)
- la production de pages Web bibliographiques devient un jeu d'enfants
- si on a besoin d'un nouveau style bibliographique, on peut en commissionner la réalisation à un informaticien (ce qui font plusieurs revues), et le distribuer ensuite à ses auteurs. Ainsi, on libère ces derniers de l'enorme perte de temps représentée par l'effort de respecter les directives de mise en page de la revue, et on assure la revue de la qualité du résultat.

Tout ça nous donne finalement la possibilité d'envisager une organisation de la connaissance sur échelle planétaire qui permette de garantir la cohérence des informations sans demander en échange de renoncer à la possibilité (ou mieux, la nécessité) de présenter aisément cette information dans les formes les plus adaptées à chaque utilisateur.

Références

[1] R. Di Cosmo. BIBTEXing au DMI. Available as `http://www.dmi.ens.fr/~dicosmo/BIBLIO/BibtexingDMI.dvi`, 1992.

[2] R. Di Cosmo. Référence brève des styles BIBTEX du DMI. Available as `http://www.dmi.ens.fr/~dicosmo/BIBLIO/DmiBiblioRefCard.dvi`, 1992.

[3] M. A. Harrison and E. V. Munson. On integrated bibliography processing. *Electron. Publ. Origin. Dissem. Des.*, 2(4) :193–209, Nov. 1989.

[4] L. Lamport. *LATEX : A Document Preparation System*. Addison-Wesley, 1986.

[5] O. Patashnik. BIBTEXing. Documentation for general BIBTEX users, 8 Feb. 1988.

[6] O. Patashnik. Designing BIBTEX styles. The part of BIBTEX's documentation that's not meant for general users, 8 Feb. 1988.