Through the mirror of a bibliography

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

Form and content. These two notions have entertained many philosophers during the course centuries in very interesting debates. Is form the only reality, since one can reach the content only through a particular form in which it is presented, or is it just an illusion, given that there is often no canonical form that is preferable to the others ?

Today, computer science can bring their modest contribution to these debates : when you have to manage an information system, the distinction between the information content and the form in which that information is presented is no longer just a philosophical question, but it can result in losses (or gains) that amount to in millions of dollars. In the branch of theoretical computer science that has been studying databases, it's been a long time that the distinction between form and content has been rediscovered, studied and mastered, despite the fact that these techniques, even though elementary, are unfortunately not always applied in the software available to the general public, due to obscure business interests that lead to despicable practices that will be discussed on another occasion.

I will try here to briefly present the reasons for the interest in computer science of this distinction, by considering some simple examples, to get to the point of describing how we can apply in the case of the management of bibliographical bases the lessons learnt.

In any computer system that manages databases, one is confronted with the following problem very much related to the distinction between form and content : how are we going to keep the information we are interested in (the content) ? This information often changes over time (the list of a company's employees, the list of publications of a researcher, the catalog of the products of a large company or store), which must be presented in different forms at different users (a bibliography in French has not at all the same look as a bibliography in English or in Italian : the style conventions are very different), and that you want to keep consistent at all times (it would be very embarrassing for a company if the prices displayed on its website were different from

those available to one of its telemarketers, or if a researcher's list of publications for the same year showed up different in two official reports from the same institution).

In early computer systems, the mistake was often made to keep several copies of the content, one in each of the forms requested by the different users. As a result of this naive organization, it became quickly difficult, if not impossible, to maintain the coherence of the information : there was no longer *one* content and several forms, but an anarchic mixture of content and forms. Since these errors were very costly, literally, it was quickly discovered that only one copy of the information content needed to be kept, to allow to update it while maintaining consistency, and then one could develop formatting programs that present this content to each user according to its needs. This is the birth of modern databases : one keeps the content in a very special form, called *logical structure*, which makes no assumption on the physical media used to store information, nor on the forms in which this information is displayed ; and one provides means to query the database by displaying the results in the most convenient form for each user.

Unfortunately, no knowledge is permanently acquired, all the more so because in the world of commercial software, so from time to time we find the same mistakes even today : without looking very far, a large company whose name I will not mention has made a great effort to put the entire catalog of its products on the Web. But in instead of doing it by the book, which meant writing a program that produces Web pages automatically from the database of the company data, they did everything by hand in a few months ; once they finished the job, they realized that it was impossible to maintain it up to date automatically (the only way to do so at a reasonable cost). I am sure that if you regularly use the Web you will have seen many examples of similar mistakes.

Let us note in passing that a similar, if not identical, phenomenon occurs with the different *format* of files used by the programs (word processing) : the author of a document is often trapped in cumbersome and poorly designed software, which forces him to mix inextricably the content (his ideas) and form (the style of presentation : e.g. some words go in bold or italics, etc.). This is a neverending source of problems every time the same content needs to be presented in another form : the same article can evolve during the course of the years and appear in different journals, or be posted on the web, but the most popular text editing software do not allow to export their content in a simple and satisfactory way, thus forcing us, the user, to waste a lot of time rewriting practically everything in the new format. Once again, this is not the place to look at the reasons, which are purely monopolistic and commercial reasons and not technological, of this widespread phenomenon, but it is necessary to recognize it in order to learn the lessons that will help us avoid falling into these traps when we are interested in the management of a bibliography.

Let's come to the subject that is close to our heart in this little note : the management of the bibliographical information. Let's try to apply the few ideas that we have to briefly exposed above : it is a matter of properly separating content and form in bibliographic information, but first of all it is important to note that there is *one* content and *several* forms to present it, which is not immediately obvious to everyone. Let's

take an example : one of my recent work is published in the following article

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

This should seem quite reasonable to anyone reading these lines, since it is the *French style* presentation of the bibliographic information [1]. However, this work is much more often cited as

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

since the *lingua franca* of computer science is English. Which of these two quotes is closer to the real content ? Consider that on my web page you will also find the Italian and Spanish versions, which are also different. The answer, for a computer scientist, is "neither" : in both cases the content is present, but inextricably mixed with presentation conventions related to the French or English language (or Italian, or Spanish etc.). Extracting the title, the name of the magazine and the authors is possible, but requires a lot of intelligence and knowledge about arbitrary conventions set by each language.

For fifteen years now, most computer scientists have been describing the content of their bibliographies by records that form a database in which finding title, author and other information is immediate, and they can be then choose the form of presentation (the *bibliographic style*) according to their needs, leaving it up to a program, BibTeX, the details of producing the proper formatting. In the case of my publication, for example, I did not handwrite neither the French version nor the English version, but only the BibTeX record that follows :

```
@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"
}
```

A few words of explanation : the line `@ARTICLE{TCS95,` says that I am defining a publication that I want to name `TCS95` for future reference, and that it is a `Article` (and not for example of a `Book`, which needs a different layout). What follows is the content itself : it is clearly stated what is the title, the year of publication etc. Even for

1. N.B. : the original article was in written in French, for a french public, hence this remark.

the author there is no prejudging of the presentation : the verb and the comma are only there as separators ; the `and` and the comma separates the authors (and it will become `et` in French, `e` in Italian etc.), while the comma is used to unambiguously identify the name and the first name when the surname or first name is composed, as is the case with the name of the author of this very article. An interesting case is that of the field `JOURNAL`, where the word `TCS` is not the full name of the journal, but an abbreviation which is reported elsewhere as follows

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

This makes it possible to imagine the implementation of a database of journal names. which ensures greater consistency in bibliographic databases, and indeed similar systems have been in use for years in some French computer labs.

A major advantage of this format is that it is *extensible*, i.e., the number of fields in a record is not fixed a priori, but may be extended at any time depending on new aspects of the content that may emerge over the years : for example, the last field `ABSTRACT-URL` is used by a style that produces references to Web pages, and it would not have not been there a few years ago. Quite simply, if a bibliographic style does not know this field, it will ignore it. Similarly, the field `DMI-CATEGORY` is used only by a class of bibliographic styles (described in the in [1, 2]) which allow to produce bibliographies that present the publications in an order that is appropriate to the the DMI's quadrennial report, but is ignored by most of the other styles. Other fields that are often found are `ABSTRACT` (which contains an abstract of the publication) and `ANNOTATION` for personal annotations.

But enough details : whoever is interested in using BIBTEXwill find plenty of information in [5, 6, 4] (these references, of course, were generated with BIBTEX itself).

What matters, is that once the information is written in this explicit structure, the final formatting will be obtained by choosing the most appropriate style, which has been programmed *once and only once* (in the previous example, the author has used the style `falpha.bst` for French and `alpha.bst` for English.). Today there is a very large collection of styles in the public domain : with or without abbreviated first names, with Web references or not, plus a style for each major IT publication (the TCS magazine uses a different one than MSCS or CACM etc.).

The advantages of this approach are very many, but let's recall some of them

— the content is described in an open format and available free of charge
— information is written once, and maintained easily, possibly with the help of advanced management systems [3]
— if different authors write in the same magazine, and all of them use this format, the bibliographies of the different articles will be perfectly homogenous and without any effort : it will be enough to tell BIBTEX to use the same format bibliography for all
— it is now possible to create bibliographic databases for general use : almost all journals or magazines or computer conferences have drawn up a list of publications in the BIBTEX format freely available on the Web, which allows you to have precise and always up to date references (a collection of more than 700,000 entries is already available in `http://wheat.uwaterloo.ca/bibliography/index.htr` for computer science)

— the production of bibliographic web pages becomes child's play
— if you need a new bibliographic style, you can commission its development to a computer scientist (which several journals do already), and then distribute it to the authors. This frees the authors from wasting a lot of their time to comply with the formatting guidelines of the journal, and ensure the journal that the result will be of constant quality.

All this finally gives us the possibility to imaing an organization of knowledge on a global scale that will ensure the consistency of the information without asking in exchange to waive the possibility (or better yet, the need) to easily present this information in the most appropriate form for each user.

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.