

## Literature and Catalogs in Electronic Form: Questions, Ideas and an Example: the IBVS

A. Holl

*Konkoly Observatory, P.O.Box 67, H-1525 Budapest, Hungary, Email:  
holl@ogyalla.konkoly.hu*

**Abstract.** While transforming astronomical journals and catalogs to electronic form, we should have in sight two questions: making it easier for the human reader to locate and comprehend information. At the same time, some of the text read by humans in the past, will be — or already is — processed by machines, and should be laid down in a different way than formerly. Information should flow more easily, but references to the origin should be kept all the way along. With the same effort, references could be checked automatically. To achieve this goal, appropriate markup should be used. Software technology has applicable ideas for this problem.

In this paper we discuss the problems of transferring old issues of astronomical journals to computerised formats, and designing formats for new material, using the example of the Information Bulletin on Variable Stars, along with experience with other journals — like the AAS CD-ROM and JAD. Some problems with machine-readable catalogs are also investigated, with ideas about improving formats (FITS) and access tools.

### 1. Introduction

While transforming astronomical journals and catalogs to electronic form, we intend to make them easier to access for the reader, and, at the same time, making them more easily processable by computers.

In this paper we discuss problems of transforming a small astronomical journal, the IAU Comm. 27 & 42 Information Bulletin on Variable Stars (IBVS) to electronic form — including about 15000 pages of previous issues, back to 1961, to the present, computer-typeset ones.

### 2. The Old Material

Rendering printed textual information to an ASCII computer file is a difficult problem. There are ambiguities in the typesetting (some old typewriters used the same character for zero and capital o, for digit one and the lowercase l character etc.), some places “redundant” characters were spared for economical reasons, math formulae, non-Latin characters and accents are common.

One could re-typeset the text in  $\text{\TeX}$ , for example, but that would be very difficult. We decided to use a format as simple as possible. We have dropped the non-Latin accents; Greek characters, math signs were replaced by

their names (as in  $\TeX$ , but without the leading backslash); for superscripts and subscripts either  $\TeX$ -like or FORTRAN-like syntax were accepted. We remove hyphenation, for the sake of simple text string searches.

Errors introduced by the Optical Character Recognition (OCR) process make the situation worse. IBVS was published mainly from camera-ready material, therefore pages were extremely heterogeneous. The OCR and primary correction was done by many different persons.

In spite of a three-pass error checking and correction process (including a spell-checker), errors still remain in the text. But what should we do with the errors originally in the text? Our accepted policy was: correct the obvious typos, spelling errors (if found), do not correct foreign spelling, nor semantical errors. We have not checked the references in the papers (except for obvious spelling errors in the names of journals).

In retrospect, we see now that we should have laid down a rule-set for the rendering in advance. It would be desirable to develop a standard, which would produce easily readable, and at the same time, computer-browsable information.

The next question is the format, in which we provide the information to the community. We have chosen plain ASCII text and PostScript images of the pages. We could have devised a simple markup for the ASCII text version, which would have enabled us, for instance, to create tables of contents automatically, or any bibliographical service provider (BSP), like ADS or SIMBAD, to process the references in the papers — we have not done this.

There is one obvious shortcoming of the ASCII text version: the figures are missing. In the final form, we will use a simple markup in the place of the missing figure, adding a brief description, if not available in the caption or in the text (e.g.: [Fig. 1.: V lightcurve for 1973]).

### 3. The New Material

For the past few years, IBVS has been typeset in  $\LaTeX$ . Source code and PostScript versions are available. Recently, we have introduced a new  $\TeX$  style file, which uses a simple markup using appropriate macro names, which enables automatic extraction of the title, author name, date information, makes possible the insertion of object, variable type (GCVS standards) and other keywords, and also abstracts. Keywords and abstracts do not appear in print, but they are part of the  $\LaTeX$  source. Macros were designed to enable the extraction of information with very simple and generally available text processing tools (i.e., Unix grep). With these new features, IBVS issues get to the Web automatically, tables of contents, object and author indices are generated automatically too. BSPs could easily process the source text (sometimes using the IBVS-specific markup, otherwise removing all  $\LaTeX$  markup completely).

Here we have to deal with the question of errors. Electronically produced material contains less misspellings, thank to the spell-checkers. But what should we do if we find a mistake in an electronic journal? Should we resist the temptation to correct such a mistake in a paper which has been already available on the Web since some time (after publication)? We adopted the following practice: we do not correct the error, but issue an erratum, which is attached to the end of a new issue (as traditionally), AND gets attached to the end (one

might use links in HTML format material) of the original issue too, and the Web-page containing the table of contents would also get a flag, notifying the reader. (This way papers could become more dynamic — one can see a journal publishing comments, discussions of papers — as in conference proceedings — attached to the original paper.)

With the references in the papers we have not done anything so far. Discussing the problem with BSPs, it would be possible to design  $\text{\LaTeX}$  macros in such a way, to help automatic reference processing.

We also have to think about the question of figures. To facilitate indexing the information content of the figures, we suggest moving most textual information from the bitmapped or preferably vector-graphic figure to the caption.

The next point to stop at is the question of tabular material. Tables in the IBVS — and in other journals available in electronic form (like the AAS CD-ROM or the Journal of Astronomical Data, also on CD) are often formatted for the human eye, and would be very difficult to read in by a program (to plot or analyze). Publishers of such journals should take care to provide tables easily processable by programs. A good example is the ADC CD-ROM series. IBVS will make available lengthy tables electronically, in machine readable ASCII text, or FITS ASCII Table form. Those tables are easily readable for humans too. The simple catalog format introduced by STARLINK should be also considered. Besides the widely used graphical or text processing tools, there are specific tools for such tables — like the Fits Table Browser by Lee E. Brozman for the ASCII FITS Tables. We have just one complaint: FTB is slow. With introducing the notion of unique (for catalog numbers) or ordered (like right ascension for many tables) variables to the FITS standard, those tools could be considerably improved.

#### 4. Formats, Media and Policy

We have decided to put all text (plain ASCII or  $\text{\LaTeX}$  source) on-line, and PostScript format issues for recent material. At the moment we do not expect to introduce PDF format, but, in the future, we might add HTML format with converting the  $\text{\LaTeX}$  sources. Our opinion is, that large volume, static material, which has a well defined user community, who regularly use the information, should be distributed on CD-ROM. So we will put old, digitized IBVS issues to a CD-ROM, in PostScript format — which we do not have storage capacity and bandwidth to provide on-line. The CD-ROM will contain IBVS issues 1-4000 and an HTML interface.

Information, which is dynamic, changeable, which is accessed casually (when a user of a BSP follows a reference), should go on-line. So we serve PostScript versions of the recent issues, and text for all. We must keep in sight that this information should be accessible to the broadest community. In consequence, we serve IBVS with different distribution methods: anonymous ftp and WWW. Readers using a public FTPMAIL server could access IBVS via e-mail too. We intend to use such HTML tags on the Web-pages, which work with all possible browsers.

We want to retain control over the textual material too — so BSPs could have it for indexing, and they could put links to the issues residing on our server

for full text. The reasons for this decision are the following: the errors in the old and new material get corrected by us, reader services are provided by us, so the best place for the material is with us. On the other hand, those investing in the project wish to retain full rights over the intellectual property. But with the Web there should be no problem with it — the interested reader, following a link, could get the material promptly, wherever it resides.

## 5. Conclusions and Remarks

Astronomical literature — old and new alike — gets on-line at a rapid pace. Besides the publishers and readers, third parties: the BSPs are concerned as well. Establishing conventions, standards would be desirable.

One can also envision — similar to software development tools and environments — publication development aids. Such tools, for example, could help check the references, whether they really point to an existing paper or not. The focus of the present FADS session is “the prospects for building customized software systems more or less automatically from existing modules”. Would it be possible to build “customized scientific papers”, automatically from existing modules? In other words, is component re-use possible for astronomical papers? I think the case of figures, tables and references should be considered.

**Acknowledgments.** The electronic IBVS project was supported by the Hungarian National Information Infrastructure Development Programme (NIIF).

## References

The IBVS homepage, URL: <http://www.konkoly.hu/IBVS/IBVS.html>

A. Holl: The electronic IBVS, Konkoly Observatory Occasional Technical Notes, No. 5, 1996