Serdica J. Computing **6** (2012), 59–66

Serdica
Journal of Computing

Bulgarian Academy of Sciences
Institute of Mathematics and Informatics

ARTICLE DISASSEMBLY—NEW WAYS TO HANDLE INFORMATION IN PUBLICATIONS

András Holl

Abstract. Articles and books – the basic publication units – could be disassembled to semantic building blocks. For scientific journal articles, such blocks include figures and tables, among others. Providing meta-data for figures and tables, and making them accessible *per se*, opens up new ways of presenting and using scientific information—like producing an image database on certain subjects, based on figures published in different journals. These meta-data, complete with copyright information, should be supplied by the publishers, who in turn might require authors to provide this information. Some examples are shown from a small astronomy journal, the *Information Bulletin on Variable Stars*.

1. Introduction. Reading a newspaper one might start (and finish) with the sports pages or the cartoons. Scientists are no different: in certain cases they might look for a figure, or check the references section first. Twenty years ago N. Negroponte contemplated the possibilities of electronic journals [1,

ACM Computing Classification System (1998): I.7.4.
Key words: library science, meta-data, electronic journals.

2]. Component re-use was mentioned by us [3] and the same idea was called decomposition of articles by K. Kroffe [4]. Some astronomical journals offer article digests for smartphones. The *Information Bulletin on Variable Stars* (IBVS) – a small, specialized astronomy journal – offers special figure and data table services. Journal articles are not monolithic entities any more—they never were in reality. Articles are to be disassembled to parts or building blocks, which will start a (somewhat) separate life from the whole paper, when these blocks could be re-assembled again to form new views, digests, or re-used in new publications, all of which could happen on demand, based on a request of a reader, using his/her personal preferences and capabilities of the reading platform.

2. Article building blocks. Let's try to list the conceptual building blocks of journal articles. The very first is

- The meta-data (present in the article at the beginning, or in the running head): title, author, publisher, publication date etc.
- Abstracts, which could be used by machines, not only humans [5].
- Next is the text itself (note that some of these building blocks could be sub-divided further).
- References (these could be re-used, as every author knows, and in ways authors might not know about). For bibliometric re-use see [6].
- Figures (which we will discuss below).
- Tables (will be discussed too).
- Data files (as additional components available only electronically, or present in the main article body as tables).
- Equations, essential information,
- etc.

3. The Case of the Figures. Figures (pictures) published by a journal constitute an informational treasure trove. Magazines have their picture databases—should not scientific journals too catalogize the figures they published? At IBVS we have a digital copy of each figure ever published in the

journal (the old figures were digitized, not only together with the articles, but separately too). Each figure is individually accessible on the web. The re-use of images is facilitated by unique identifiers, and figure meta-data, as keywords, object keywords and captions. Article meta-data (author, for instance) are inherited too. (One could imagine situations where figures have different authors—this is not the case in astronomy, where they are usually attributed to, and created by, the authors of the paper.) The presence of figure meta-data facilitates figure search. At IBVS readers can search for frequently used figure types (finding chart, light curve) for a specific object. Figures can be accessed without the need of downloading the whole article—but with links to it, naturally. These images could be embedded in other services – and re-used that way – too. The WEBDA database [7] has some re-published figures from IBVS. The readers of the electronic version could choose different file formats for download (JPEG, PostScript or occasionally GIF).

Figures themselves could be decomposed: they consist of data ($x - y$ number pairs for a scatterplot), maybe a background image, labels and drawing instructions (program code that could produce the figure using the data provided and govern the placement of labels, the axes and tick marks, colors and line weights, etc.). Data, image, labels and instructions – and whatever else is needed – can be embedded in a single XML file. Such presentation of a figure would not only enable re-use (one could use a figure from the literature with some new data points added, for example), but more. In the electronic edition the provenance of each point plotted could be examined by moving the cursor on top of it. Also, such hyper-vector graphics would enable different formatting of the same figure for different media: paper, web, mobile phone, not only with using proper line weights, but with omitting less important information from a crowded graph when viewed on the small screen of a phone.

IBVS offers a third-party visualization and CIS (Celestial Information System, meaning software similar to Geographical Information Systems, but showing the sky) tool for some figures: the Aladin of the CDS, Strasbourg [8]. With Aladin readers can compare the figures with standard sky surveys, over-plot catalog data, or measure distances.

The number of figures stored at IBVS presently exceeds 11000, including some animated GIFs.

4. The Case of Tables and Data Files. In science, tables (together with the already discussed figures) often hold the most important information in an article. K. Kroffe mentioned putting more data behind the figures for the

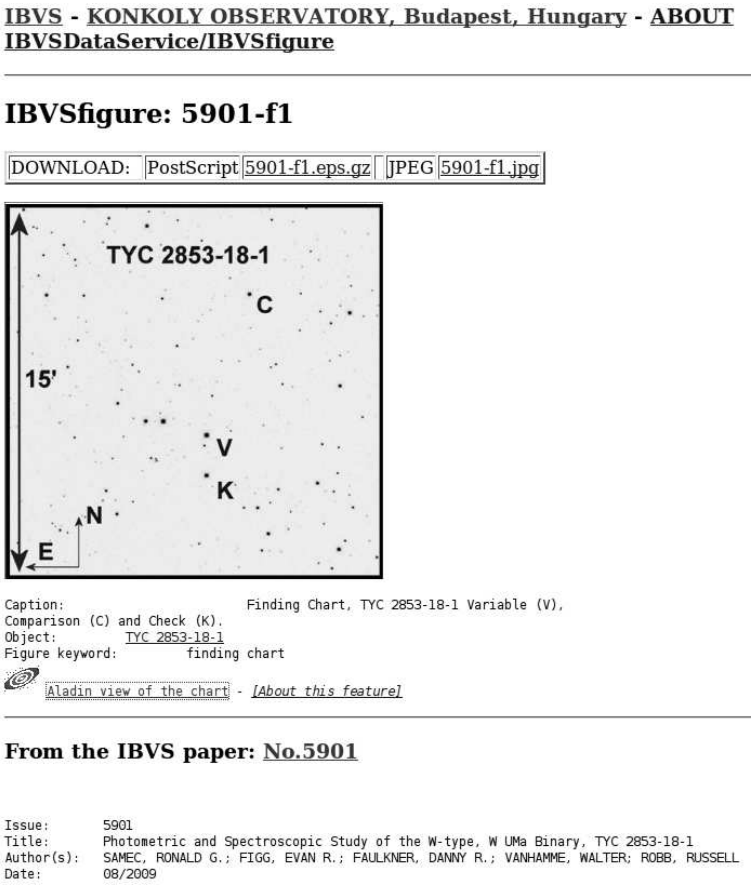

Fig. 1. The IBVSfigure service presents a figure with meta-data, and further options

American Astronomical Society journals published by IoP [4], and the same idea is practiced at IBVS too. Data files containing information which appeared in tables or in figures often accompany the articles. In variable star astronomy time-series photometry is one of the most commonly used data types, which often appears in the form of light curve plots. IBVS encourages authors to provide such data in separate files, if not present in tables already. Tables have a basically machine-readable structure, but printed tables are often formatted for the human eye, and could not be interpreted easily by computers. At IBVS machine-readable electronic versions of tables are used. Really machine-readable tables in astronomy should be in VOTable format—although these are, conversely, difficult for people to read. Tables need to be presentable both for humans and programs,

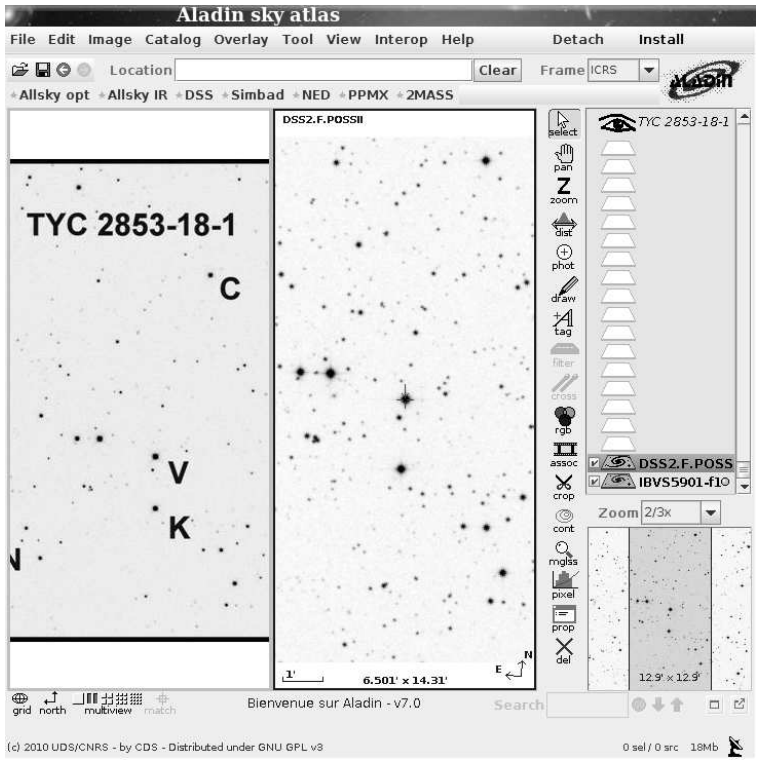

Fig. 2. The Aladin visualization of the chart from IBVS. In the left-hand panel the figure from IBVS is shown, the right-hand panel contains the interactive sky chart of the same area

so we need tools for rendering VOTables.

At IBVS electronic tables (data files), just like figures, are accessible and searchable on their own, and come with their unique identifiers. (DataCite [9] offers DOIs for data files—IBVS uses local identifiers, but these could be converted to global identifiers.) They have their own meta-data, including keywords and object keywords. Where available, a choice of different formats (LaTeX, TXT, occasionally VOTable [10]) are offered to the reader. LaTeX tables might contain hyperlinks.

IBVS uses CDS Aladin for the visualization of certain types of tabular material [11]. The identifiers of tables are reported to CDS and ADS along with the standard article-level meta-data, making these data files directly accessible from the ADS Search service. There are more than 1300 data files stored at IBVS.

Data files pose a problem for long term preservation. Standard or simple file formats should be used: plain text or, in astronomy, FITS [12] and VOTable XML.

5. Essential information. Sometimes the essence of a journal article is a single number (or a few numbers). In astronomy, in the field of variable star research (the subject area of IBVS), such numbers could be the periods of eclipsing binaries in a discovery note. In archeology the age of a find is such an important quantity. It is crucial to expose these numbers to harvesters for the creation of the semantic web. One solution for making these important quantities machine readable is the use of small – even single-row – tables. IBVS already employs standardized forms – which use LaTeX tables – for discovery notes. Another way could be using the technologies of the semantic web, like OAI-ORE [13].

IBVS regularly publishes articles with minimum or maximum times of variables. These papers consist of only a few lines or paragraphs of text, describing the observational methods, and then a lengthy table, containing hundreds or thousands of lines. The table is typeset in LaTeX, but we produce machine-readable plain-text versions too. These articles are the most often cited ones, where the data is used in "normal" articles on a single object (or few objects). Several databases use such data from IBVS [14]—automatic data transfer from IBVS articles to such databases seems inevitable. One might ask: why publish such tables in journals? Should not they go directly to databases? Maybe. But the problem of data citation needs to be solved first.

6. Article re-assembly and re-use, copyright issues. Disassembled article building blocks could be re-assembled again. One could create personalized journals, mashups, databases. For this building blocks, and their relations should be described—OAI-ORE provides a way for doing that. An example description of an astronomy article is available [15].

Copyright issues need to be addressed before article components could be re-used widely. While we do not think that Creative Commons (or GNU GPL) licenses are appropriate for the whole scientific journal article, they could be used for some of the article building blocks. For the figures CC BY-NC (attribution and non-commercial) or BY-NC-SA (with the added share alike condition) seems to be appropriate. Declaring the copyright model is not enough: we need to provide protocols to communicate the copyrights together with the other metadata and the article component itself. Autonomous software agents should not

only be capable of discovering the resource (e.g., a figure in the article), but should be able to ascertain that it is lawfully re-usable.

7. Conclusion. There are deeper levels of meta-information in scientific publications than commonly used (volume or article level). Exposition of content is widely practiced in library science. We demonstrated that meta-information on figures and tables should be provided by the publishers, and figures and tables should be made accessible, organized into databases, allowing the production of a richer scientific e-journal environment for the readers. The Information Bulletin on Variable Stars does offer some innovative features for some article blocks already.

REFERENCES

[1] Wright K. The Road to the Global Village. *Scientific American*, **262** (1990), No 3, 83–94.

[2] Negroponte N. P. Products and Services for Computer Networks. *Scientific American*, **265** (1991), No 3, 76–83.

[3] Holl A. Journals on the web – more than text. iAstro/IDHA workshop talk. `http://www.konkoly.hu/staff/holl/journals.html`, 2002

[4] Isaksson E., J. Lagerstrom, A. Holl, N. Bawdekar. Library and Information Services in Astronomy VI: 21st Century Astronomy Librarianship, From New Ideas to Action. In: Proceedings of the ASP Conf., Pune, Maharashtra, India, Feb. 14-17, 2010 (Eds E. Isaksson, J. Lagerstrom, A. Holl, N. Bawdekar.), ASP Conf. Ser., Vol. **433**, Astronomical Society of the Pacific, San Francisco, 2010, 355–359.

[5] Henneken E. A., A. Accomazzi, M. J. Kurtz et al. Exploring the Astronomy Literature Landscape. In: Proceedings of the ASP Conf. ADASS XVIII (Eds D. A Bohlender, D. Durand, P. Dowler), ASP Conf. Ser., Vol. **411**, Astronomical Society of the Pacific, San Francisco, 2009, 384–387.

[6] Kurtz M. Second Order Knowledge: Information Retrieval in the Terabyte Era. In: Astronomy from Large Databases II, (Eds A. Heck, F. Murtagh), ESOC 43, European Southern Observatory, Garching, 1992, 85–85.

[7] WEBDA, http://www.univie.ac.at/webda/

[8] CDS Aladin. http://aladin.u-strasbg.fr/aladin.gml

[9] DataCite, http://datacite.org/

[10] VOTable documentation. http://cdsweb.u-strasbg.fr/doc-cds/VOTable/

[11] Holl A. IBVS and the data from robotic observatories. *Astronomische Nachrichten*, **325** (2004), 610–612.

[12] Flexible Image Transport System (FITS). http://fits.gsfc.nasa.gov/

[13] Lagoze C., H. Van de Sompel, M. L. Nelson, S. Warner, R. Sanderson, P. Johnston. Object Re-use and Exchange: A Resource-Centric Approach. (Cite as: arXiv.0804.2273v1 [cs.DL]). http://arxiv.org/abs/0804.2273v1, 2008

[14] Holl A. Observations and Publications in the VO: is the VO only for Big Science? In: Proceedings of the ASP Conf. Cambridge, Ma, 2006 (Eds S. Ricketts, C. Birdie, E Isaksson), Library and Information Services in Astronomy V, ASP Conf. Ser., Vol. **377**, Astronomical Society of the Pacific, San Francisco, 2007, 47–52.

[15] Reynolds D. Astronomy Data Sample Article https://wiki.library.jhu.edu/display/DATAPUB/OAI-ORE+Model+for+Sample+Article+1

András Holl
Konkoly Observatory
of the Hungarian Academy of Sciences
Library of the Hungarian Academy of Sciences
Budapest, Hungary
e-mail: holl@konkoly.hu

Received October 31, 2011
Final Accepted March 9, 2012