Astronomical Data Analysis Software and Systems VIII
ASP Conference Series, Vol. 172, 1999
D. M. Mehringer, R. L. Plante, and D. A. Roberts, eds.

The ADS Bibliographic Reference Resolver

Alberto Accomazzi, Guenther Eichhorn, Michael J. Kurtz, Carolyn S. Grant, Stephen S. Murray

Smithsonian Astrophysical Observatory, Cambridge, MA 02138, USA

Abstract. An increasingly larger number of publishers and institutions are using the NASA Astrophysics Data System (ADS) to verify the existence and availability of references published in the astronomical literature. In this paper we discuss the tools and utilities that the ADS is developing to provide the capability to automatically parse, identify, and verify the existence and correctness of citations appearing in astronomical publications. Details of the implementation are presented and some preliminary results from the use of the resolver are shown.

1. Purpose of the Reference Resolver

The capability to interactively access publications cited in scientific papers is of undisputed importance to researchers and librarians, greatly increasing their efficiency and quality of work. Thus it is increasingly important for publishers and data providers to be able to automatically verify the syntactic and semantic correctness of the references listed in their online documents as a way to improve the overall quality and accuracy of the information they are providing. Verifying the existence of a reference through a well-established online database such as the ADS also allows them to publish documents containing hyper-links pointing to stable, unique URIs. Additionally, when verifying references via a "smart" resolver (such as the one described in this paper), they contribute new entries to the citation (or "forward reference") database maintained by the provider of the resolver, allowing the creation of citation links pointing back to their site.

2. Implementation

Generally stated, the problem of resolving a reference to a paper published in the literature can be broken down into the following steps: reference parsing, i.e., identifying the different tokens listed in the reference entry; reference identification, i.e., using the parsed tokens to generate a unique identifier for the reference; reference verification, i.e., querying a database to verify the existence of the reference. The degree of success with which each of these steps is performed is dependent upon the outcome of the preceding steps.

In order to provide a flexible and portable way to represent the dataset being processed, the input data is transformed from a sequence of strings into data structures which hold all the information accumulated during the resolving

process. Each of the resolving steps listed above simply takes the data structure generated by the previous step as input, modifies it (typically by adding records to it), and then feeds it to the following processing step. To increase the overall resolving success rate, any ancillary information (or "hints") collected by each of the resolving steps is also passed on to the subsequent task. A priori knowledge about the format or origin of the dataset being resolved can be specified via a configuration file, which sets default values for some of the parameters controlling the processing performed by the reference resolver.

Once all the resolving steps have been carried out on a reference, the generated data structure is serialized as an instance of an Extensible Mark-up Language (XML) document. XML was chosen as the output format because of its flexibility in representing complex data structures and for its increasingly significant role as a standard data exchange format.

2.1. Reference Parsing

The first task performed by the resolver is parsing of the input reference in order to identify the fields that compose the bibliographic reference string. The reference listing conventions adopted by all the major astronomical journals since 1991 instruct authors to format entries in slightly different ways depending on the nature of the publication (see example, §2.2.). Other journals or conference proceedings series may adopt different conventions and syntaxes for listing references, often making it impossible to correctly break down the items composing a reference without a priori knowledge. In addition, to accommodate the needs of editors and publishers who intend to integrate the reference resolver in their electronic publication process, we had to take into account the fact that the input reference strings to be parsed could possibly be formatted and encoded in a few different ways (the most typical ones being LaTeX or SGML), and could make use of publisher-specific macros or entity definitions. Therefore, our implementation of the parser is driven by parameters specified in configuration files which are style- and format-specific.

2.2. Reference Identification

Identifying a reference is the activity of mapping the parsed fields extracted from the reference entry into a bibliographic code (Schmitz et al. 1995), the standard format used by the ADS to uniquely identify bibliographic entities. The set of fields resulting from the successful parsing of a typical reference may include: Authors, Publication Year, Journal or Conference Series name, Publication Volume, Page Number, Conference or Book Title, Editor, and Publisher. In order to be able to successfully resolve references containing journal abbreviations and macro names, the resolver uses external tables to map such names into the journal abbreviations used in bibliographic codes. In order to minimize errors in interpreting these strings, the resolver makes use of information supplied by the user via the input configuration file and any other hints collected from the parsing stage. For instance, by noting that the input string is in LaTeX format, that the journal name is specified as a macro, and that the macro style has been declared to be "AAS" in the configuration file, the resolver will match the journal strings against the ones listed in the AAS LaTeX macro package.

Example Examples of output records from the ADS reference resolver. The input reference strings appear in each record within the `<input>` tags.

```
<record>
<authors>
<item>Drinkwater, M. J.</item>
<item>Wood, P. R.</item>
</authors>
<bibcode>1985mlrg.proc..257D</bibcode>
<hints isconf="1" status="VERIFIED" istex="1" parsed="2" />
<input>Drinkwater, M.J., \&amp; Wood, P.R. 1985, in Mass Loss from Red
Giants, ed. M. Morris \&amp; B. Zuckerman, (Dordrecht: Reidel), 257</input>
<journal>in Mass Loss from Red Giants, ed. M. Morris &amp; B. Zuckerman,
(Dordrecht: Reidel)</journal>
<matches>
<item bibcode="1985mlrg.proc..257D" score="1" />
</matches>
<page>257</page>
<year>1985</year>
</record>

<record>
<authors>
<item>Pijpers, F. P.</item>
<item>Hearn, A. G.</item>
</authors>
<bibcode>1989A&amp;amp;A...209..298P</bibcode>
<hints errmsg="error in input reference, does not exist"
status="IDENTIFIED" errcode="321" istex="1" ismacro="1" parsed="3" />
<input>Pijpers, F.P., \&amp; Hearn, A.G. 1989, \aap, 209, 298</input>
<journal>\aap</journal>
<matches>
<item bibcode="1989A&amp;amp;A...209..198P" score="0.9" />
</matches>
<page>298</page>
<volume>209</volume>
<year>1989</year>
</record>

<record>
<authors>
<item>Allard, F.</item>
</authors>
<bibcode>1997ARA&amp;amp;A..35..137A</bibcode>
<hints errmsg="partial input reference, exact match not possible"
inpress="1" status="VERIFIED" errcode="311" istex="1" authtrim="1"
ismacro="1" parsed="2" />
<input>Allard, F., et al.  1997, \araa, in press</input>
<journal>\araa</journal>
<matches>
<item bibcode="1997ARA&amp;amp;A..35..137A" score="0.55" />
</matches>
<year>1997</year>
</record>
```

2.3. Reference Verification

Once a bibliographic code has been generated by the reference identification step, the following verification steps are attempted:

- Verify that the ADS contains an entry that matches the input bibliographic code and list of authors. If an exact match is found, return it to the user.
- If the input reference is incomplete because not all the necessary information required to compute an exact bibliographic code has been specified (e.g., a paper which is in press), attempt partial matches constraining the results to match the partial information supplied, and return a list of plausible matches if any were found.
- If the input reference appears in a publication source and date range for which the ADS has complete coverage, then something in the input reference is incorrect. Under these circumstances the resolver attempts several

fuzzy matches between the input reference and the ADS databases, returning any records with high correlation scores.

- If none of the previous steps has been successful in identifying a reason for the failed match, the resolver can be asked to attempt to find the best match of the input reference against the ADS databases, by iteratively relaxing the search constraints, and then return any possible matches.

3. Conclusions

We have presented the design and implementation of a reference resolving service which is currently being built by the ADS group. The resolver, currently to be considered a beta release of what is to become the standard server, already allows sophisticated matching of a variety of references against the contents of the ADS database, offering valuable feedback whenever a particular reference cannot be identified and matched. Recent use of the service has reported a success rate of 80% when matching references listed in an ASP Conference Proceedings volume.

Planned enhancements to the capabilities currently implemented include: support for a "strict" mode which would flag any error in the formatting of the input reference, even if its resolution can be completed, and a "loose" mode which would be more lenient in the verification process; providing a stable client library which can be used to seamlessly integrate reference resolution into different electronic publishing environments; and support for additional output formats, in particular the Astronomical Mark-up Language or AML (Guillaume & Murtagh 1999) and all the custom formats currently generated by the ADS search interface.

For more information on the status of the ADS reference resolver, please visit the URL `http://adswww.harvard.edu/pubs/resolver`.

References

Guillaume, D. & Murtagh, F. 1999, this volume, 278

Schmitz, M., Helou, G., Dubois, P., LaGue, C., Madore, B., Corwin, H. G., & Lesteven, S. 1995, in Information & On-line Data in Astronomy, ed. D. Egret & M. A. Albrecht (Dordrecht: Kluwer), 271