*Library and Information Services in Astronomy III*
*ASP Conference Series, Vol. 153, 1998*
*U. Grothkopf, H. Andernach, S. Stevens-Rayburn, and M. Gomez (eds.)*

# Addressing the Heterogeneity of Subject Indexing in the ADS Databases

David S. Dubin

*Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, USA*

**Abstract.** A drawback of the current document representation scheme in the ADS abstract service is its heterogeneous subject indexing. Several related but inconsistent indexing languages are represented in ADS. A method of reconciling some indexing inconsistencies is described. Using lexical similarity alone, one out of six ADS descriptors can be automatically mapped to some other descriptor. Analysis of postings data can direct administrators to those mergings it is most important to check for errors.

## 1. The ADS System

The Astronomy and Astrophysics Abstract Service of the NASA-funded Astrophysics Data System (ADS) (Eichhorn et al. 1998) serves some 10,000 active astronomers worldwide. ADS provides access via the World Wide Web to over a million abstracts in the areas of astronomy and astrophysics, instrumentation, physics and geophysics. Users of the abstract service can search by title, author, publication date, SIMBAD/NED/LPI object name, and in the texts of the abstracts themselves. Over 40,000 of the abstracts include links to scanned images of the full journal articles.

Records in ADS include subject descriptors, but the default search interface includes no mechanism for searching subject descriptors alone. Instead, users may search on terms included in either the subject descriptor field or the abstract field. On ADS the subject descriptor field is called the “keyword field” and descriptors are referred to as “keywords.” However, the contents of the keyword field include a variety of descriptor types, including index terms, precoordinated subject headings, and journal-specific keywords. The ADS administrators decided to merge the abstract and keyword indexes to help overcome limitations of this heterogeneous subject indexing. Users of ADS may still search the keyword field separately through a “legacy search form.”

NASA’s Scientific and Technical Information group provided ADS with abstracts for articles published between 1975 and 1995. Those abstracts had index terms assigned from the NASA Thesaurus. Since mid-1995, ADS has received the majority of abstracts directly from journals. Abstracts from journals have descriptors assigned from one of several related but inconsistent sets of headings. As ADS has accumulated records from more different sources, other controlled vocabularies have become included. For example, records obtained from the Library of Congress (LC) are indexed with LC subject headings.

Researchers at the University of Illinois were invited to investigate methods of merging the indexing languages in use in ADS, and any other methods of dealing with the heterogeneity. Earlier, a method for automatically associating terms with each other through factor analysis had been proposed (Kurtz 1993; Ossorio 1966). This paper reports results of preliminary experiments for a method based on lexical similarity.

## 2. Research in Vocabulary Switching and Merging

Vocabulary control is a standardization of the way concepts are named in an index. Controlled descriptors improve information retrieval systems in several ways:

1. They provide additional access points for documents in the database. An author may use one particular term to refer to a concept, but an indexer may apply a different one. The document can then be retrieved on a search for either term.

2. They overcome variation in natural language by providing standardized labels for concepts. By searching on a controlled descriptor, the user of a database need not think of all possible ways in which authors may have referred to the concept of interest.

3. They often provide indexers and searchers with links between terms that represent relationships like association, genus/species, or whole/part. This system of linkages is referred to as a *syndetic structure*.

4. They allow indexers to identify the most central or important concepts in a text. Users of a database can then restrict their searches to those specific concepts, rather than all that the author may have mentioned.

When documents in a database are indexed with one of several different systems, the benefits of controlled vocabulary are lost. For example, searching ADS on the keyword **Wolf-Rayet Stars** returns a different set of documents than a search on **Stars: Wolf-Rayet**. To address these inconsistencies, the ADS administrators disabled the ability to search keywords alone through the default search interface. As a result, one cannot search on concepts identified as most important by an author, editor, or indexer.

Efforts to reconcile different indexing languages go back at least thirty years (Wall & Barnes 1969). The literature of this problem includes treatments of the general problem of compatibility between indexing languages (Svenonius 1983; Lancaster & Smith 1983; Dahlberg 1981; Dahlberg 1983; Rada 1990), methods for achieving a merging or synthesis (Smith 1974; Klingbiel 1985; Rada 1987; Smith 1992; Amba 1996; Sintichakis & Constantopoulos 1997), and detailed descriptions of relationships found to hold between descriptors in different systems (Block 1978). Applications have been reported for indexing languages in astronomy (Silvester & Klingbiel 1993) and other disciplines (Chaplan 1995; Niehoff & Mack 1985).

## 3. Defining and Addressing the Problem

The most obvious problem with inconsistent indexing languages is that the same concept may have more than one label. But the mapping between terms in two different systems of descriptors is not necessarily one-to-one: a term may have no counterpart in another system, or there may be more than one candidate. Furthermore, two terms may have a relationship other than synonymy or exact correspondence. Lancaster & Smith (1983) discuss several possible relationships:

- There may be an exact match between descriptors (e.g., **Hydrodynamics** and **Hydrodynamics**).
- Descriptors may have slightly different spellings (e.g., **Color** and **Colour**).
- Descriptors may differ in punctuation or word order (e.g., **Low Mass Stars** and **Stars: low-mass**).
- A descriptor may have a synonymous counterpart (e.g., **Andromeda Galaxy** and **Galaxies: Individual Messier Number: M31**).
- A descriptor's closest counterpart may be a broader or narrower term (e.g., **Cosmology: Cosmic Microwave Background** and **Cosmology**).
- A precoordinated descriptor may map to more than one counterpart (e.g., **Microwave Background Radiation** to **Background Radiation** and **Microwaves**).
- A descriptor may map to two or more counterparts through "semantic factoring" (e.g., **Thermometer** to **Temperature**, **Measurement**, and **Instrumentation**).

Methods of merging controlled vocabularies (or automatically identifying correspondences between them) employ several different sources of evidence:

- Methods usually employ some kind of lexical normalization or matching on the basis of similar spelling.
- Some approaches use the syndetic structure indexing languages to suggest likely correspondences. For example, if no match for a term can be found, an algorithm may try to find a match for a superordinate or broader term.
- Some approaches include the analysis of documents jointly indexed under two different systems. The goal is to look for pairs of terms from different indexing languages that are consistently applied to the same documents.
- Another source of evidence is the judgment of human indexers or domain experts. Human judgments can be used in semi-automated approaches to make final decisions, or to evaluate the success of fully automated systems.

In attempting to improve subject access for ADS, several solution goals were deemed desirable. Correspondences identified by the system should afford interpretation and verification by human experts. Evaluation of the system's success must ultimately be tested in the use of the system by ordinary searchers, but it is important for the ADS administrators to see and understand which descriptors are being connected and why[1]. A solution ought to provide insight into the nature and scope of the problem it solves: ADS administrators lacked direct evidence of how seriously the indexing problem was interfering with effective searching. Finally, a simple, computationally inexpensive solution is desirable, even if only as a basis of comparison for more sophisticated methods.

## 4. A Lexical Matching Method

A simple approach based on lexical similarity was adopted, based on success reported in a previous study (Sintichakis & Constantopoulos 1997). Each subject descriptor was converted to a "lexical signature" according to the following algorithm:

1. Common stop words (such as "and," "of," and "the") are removed.
2. All punctuation marks are removed.
3. The Porter stemming algorithm (Porter 1980) is applied to remove suffixes.
4. The remaining word stems are permuted into alphabetical order and concatenated.

For example, at the time this study was conducted, ADS contained four variations of the heading **radiation mechanisms: non-thermal** (including and excluding both the hyphen and the final 's' in 'mechanisms'). The lexical signature for each of the four is "mechannonradythermal." Each such signature represents a cluster of descriptors, grouped together based on similar spelling. This approach to mapping between vocabularies is simple to implement, computationally inexpensive, and produces clusters that can be inspected for validity.

One out of every six descriptors in ADS maps to at least one other lexically similar term. However, many of these appear to be spelling errors that occur in few documents. By linking the members of these lexical clusters to their postings counts in ADS, a limited picture emerges of the method's potential impact on search effectiveness.

Imagine a version of ADS in which terms in a keyword query are automatically expanded to include all lexical variants of the term. Suppose that a user of the system searches on one term at a time. Make the further (conservative) assumption that the user invariably searches on the most popular variant (i.e., the one with the highest number of postings). Distributions of additional retrieved documents and of an impact factor under these assumptions are presented in Figures 1 and 2. These distributions are over every lexical cluster (i.e. one sixth of the ADS indexing vocabulary).

Figure 1 shows a distribution of additional documents (after a base ten logarithmic transformation), under the assumptions outlined in the previous paragraph. As can be seen in the figure, expanding to all lexical variations

[1] Difficult-to-interpret methods include factor analytic approaches, where related terms are mapped near to each other in a multidimensional space (Dumais et al. 1988).

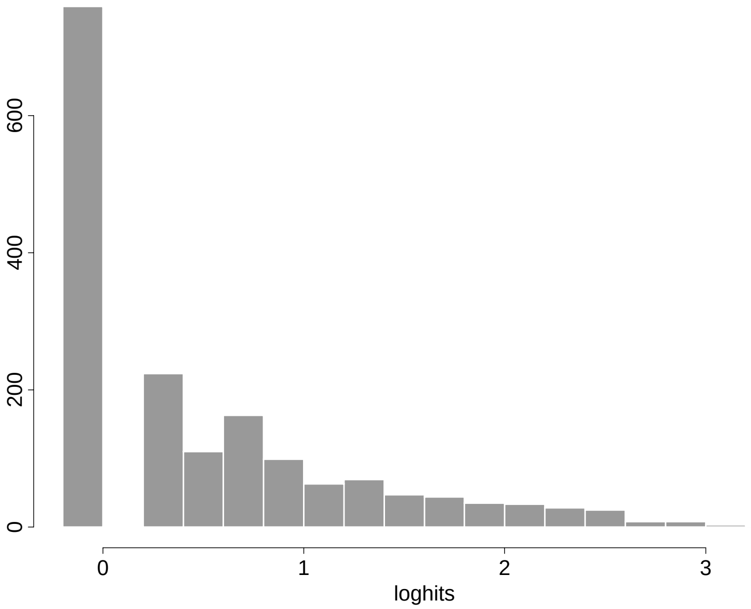

Figure 1. Distribution of the number of additional documents

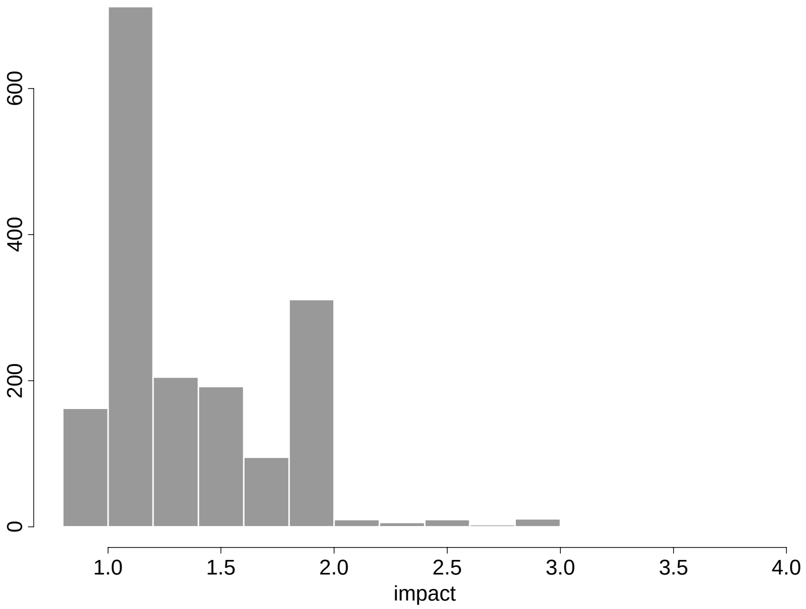

Figure 2. Distribution of the impact measure

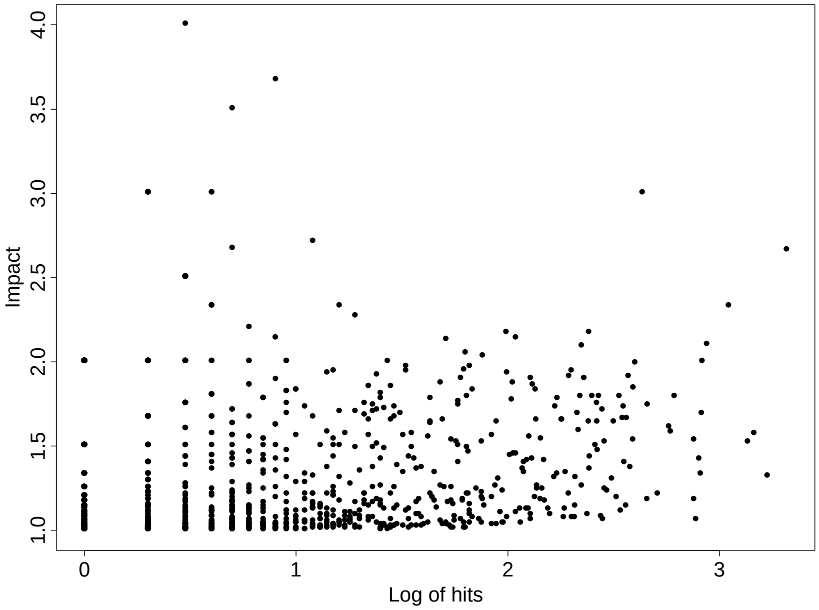

Figure 3. Plot of hits vs. impact

could result in hundreds of additional retrieved documents. Typically, however, such expansion will only produce one or two additional documents. Figure 2 expresses the same data as a ratio. For example, if expansion were to double the number of retrieved documents, then the impact factor would be 2.0. The figure shows that expansion of some terms will increase the number of hits by a factor of three. However, the median value of that distribution is 1.2 (i.e., a twenty percent increase over the most common variant).

The data in figures 1 and 2 give no information on whether any of the mappings are correct or erroneous. An evaluation of the overall success of the method requires further analysis. However, the existing data alone can direct ADS administrators' attention to those mergings which, if in error, are likely to have the most serious detrimental effect on search precision. Figure 3 shows a scatter plot of the hits measure against the impact measure. Term clusters represented by points at the top and right sides of the graph are those that will have largest effect if term mappings are in error.

## 5. Future Work

The lexical matching method holds the most promise for terms applied to relatively few documents (i.e., where three or four additional documents represents a significant increase in recall). Future work will compare more sophisticated and subtle term mapping methods to the lexical matching method. Current research includes the analysis of jointly indexed documents using a spreading activation model (Lee 1998).

## References

Amba, S. 1996, in SIGIR 96: Proc. 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, H. P. Frei, D. Harman, P. Schauble, & R. Wilkinson, eds. (New York: Association for Computing Machinery), 181

Block, M. L. 1978, Analysis of Existing Controlled Vocabularies, (San Francisco: Women's Educational Equity Communications Network)

Chaplan, M. A. 1995, The Library Quarterly, 65, 39

Dahlberg, I. 1981, International Classification, 8, 86

Dahlberg, I. 1983, International Classification, 10, 5

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. 1988, in Human Factors in Computing Systems: Proceedings of CHI'88, (Baltimore: Association for Computing Machinery), 281

Eichhorn, G., Accomazzi, A., Grant, C. S., Kurtz, M. J., & Murray, S. S. 1998, in Astronomical Data Analysis Software and Systems VII: ASP Conf. Ser. 145, R. Albrecht, R. N. Hook, & H. A. Bushouse, eds. (San Francisco: Astronomical Society of the Pacific), 378

Klingbiel, P. H. 1985, Information Processing and Management, 21, 113

Kurtz, M. J. 1993, in Intelligent Information Retrieval: the Case of Astronomy and Related Space Sciences, A. Heck & F. Murtagh, eds. (Dordrecht: Kluwer), Astrophysics and Space Science Library, Vol. 182, 21

Lancaster, F. W. & Smith, L. C. 1983, Compatibility issues affecting information systems and services, (Paris: United Nations Educational, Scientific, and Cultural Organization)

Lee, J. 1998, A Theory of Spreading Activation for Database Merging, (Urbana-Champaign, IL: GSLIS, University of Illinois), Unpublished report

Niehoff, R. & Mack, G. 1985, International Classification, 12, 2

Ossorio, P. 1966, Multivariate Behavioral Research, 1, 479

Porter, M. F. 1980, Program, 14, 130

Rada, R. 1987, International Classification, 14, 63

Rada, R. 1990, International Classification, 17, 158

Silvester, J. P. & Klingbiel, P. H. 1993, Information Processing and Management, 29, 47

Sintichakis, M. & Constantopoulos, P. 1997, in SIGIR 97: Proc. 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, N. Belkin, A. D. Narasimhalu, & P. Willett, eds. (New York: Association for Computing Machinery), 129

Smith, L. C. 1974, Journal of the American Society for Information Science, 25, 343

Smith, L. C. 1992, in Classification Research for Knowledge Representation and Organization : Proc. 5th International Study Conference on Classification Research, N. J. Williamson & M. Hudon, eds. (Amsterdam: Elsevier), 337

Svenonius, E. 1983, International Classification, 10, 2

Wall, E. & Barnes, J. M. 1969, Intersystem compatibility and convertibility of subject vocabularies, Technical Report 1582-100-TR-5, (Philadelphia: Auerbach Corporation)