A Tony Thomas-Inspired Guide to INSPIRE

Heath B. O'Connell

Fermilab, Batavia IL 60510, USA
hoc@fnal.gov

Abstract.
The SPIRES database was created in the late 1960s to catalogue the high energy physics preprints received by the SLAC Library. In the early 1990s it became the first database on the web and the first website outside of Europe. Although indispensible to the HEP community, its aging software infrastructure is becoming a serious liability. In a joint project involving CERN, DESY, Fermilab and SLAC, a new database, INSPIRE, is being created to replace SPIRES using CERN's modern, open-source Invenio database software. INSPIRE will maintain the content and functionality of SPIRES plus many new features. I describe this evolution from the birth of SPIRES to the current day, noting that the career of Tony Thomas spans this timeline.

Keywords: SPIRES, INSPIRE, database, library, literature
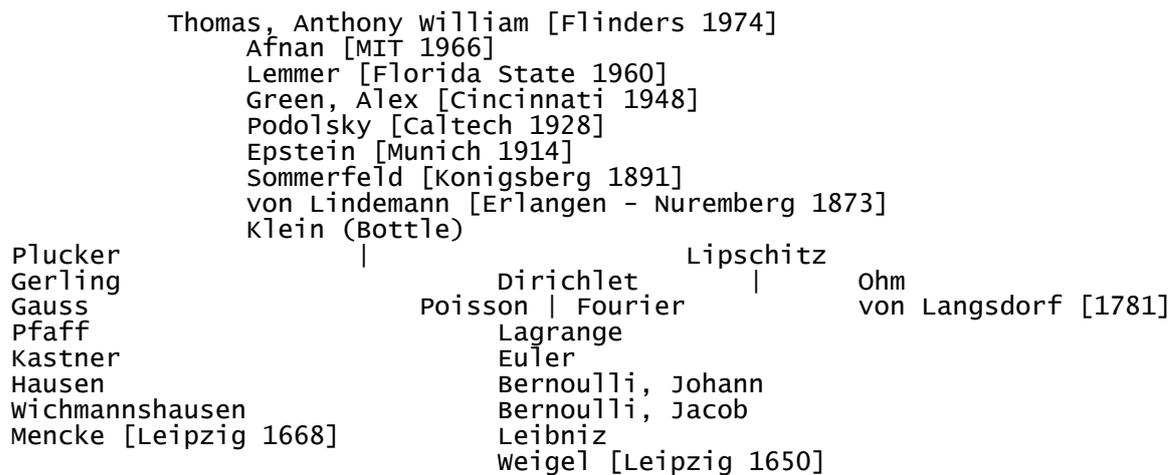**PACS:** 89.20.Hh

INTRODUCTION

As the most fundamental of sciences, Physics has traditionally asked the Big Questions. The role of the SPIRES database is to provide help in answering these questions by making it as easy as possible to discover previous work on the subject of interest. Although the main SPIRES database, HEP, is a database for literature (both published and unpublished), Physics as an endeavour involves people (listed in the Hepnames database), places (Institutions), people meeting at various places (Conference) and a recognition that at the end of the day, one has to put food on the table (Jobs).

How did we get here?

The power of this suite of interactive databases allows one to investigate questions such as "How did we get here?" The HEPNAMES database includes (where the information is known) each person's Ph.D. supervisor (in the U.S. one would say "advisor"). From this, it is relatively easy to indulge in some academic genealogy, and we can trace Tony Thomas' impressive pedigree all the way back to the 17th Century (Fig. 1). It can be seen that this goes beyond a family tree-level graph by including loops.

How did it all begin?

Although the SLAC Library began cataloguing preprints in the SPIRES database in 1968, the version of SPIRES on the web dates back traditionally only to 1974 when

```
          Thomas, Anthony William [Flinders 1974]
                Afnan [MIT 1966]
                Lemmer [Florida State 1960]
                Green, Alex [Cincinnati 1948]
                Podolsky [Caltech 1928]
                Epstein [Munich 1914]
                Sommerfeld [Konigsberg 1891]
                von Lindemann [Erlangen - Nuremberg 1873]
                Klein (Bottle)
Plucker                              |                Lipschitz
Gerling                         Dirichlet      |      Ohm
Gauss                     Poisson | Fourier         von Langsdorf [1781]
Pfaff                           Lagrange
Kastner                         Euler
Hausen                          Bernoulli, Johann
Wichmannshausen                 Bernoulli, Jacob
Mencke [Leipzig 1668]           Leibniz
                                Weigel [Leipzig 1650]
```

FIGURE 1. The academic ancestry of Tony Thomas.

the collaboration between SLAC and DESY began [1]. Although one might think that this would be sufficient for the needs of modern researchers, many older papers are still highly relevant; a 1951 paper by Schwinger [2] received over 100 citations in 2009. We have therefore worked to ensure that older papers are well represented in SPIRES first by including the papers from 1968-1974 that were stored in an old version of the database, then by working closely with publishers such as the American Physical Society, who enabled us to ensure that all papers from *Physical Review C* and *D* were systematically included. Finally, we encourage submissions from our users, the world-wide HEP community, telling us about missing papers and provide an easy-to-use form for this purpose. Where possible we link to the full-text of these articles, now widely available on publishers' servers or as preprint scans at various HEP laboratories such as KEK in Japan. Thanks to this effort, you can find Tony Thomas' first published paper in HEP[1] (see Fig. 2).

What are the fundamental building blocks?

Modern web tools allow one to visualise data in new ways and this could be done with SPIRES metadata.[2] To take a simple example, it would be possible to do a search and then present the words from the titles of the papers in a "word cloud." Fig. 3 shows the "title cloud" of Tony Thomas 600+ papers, the larger the word, the more frequently it appears in the titles.

---

[1] I thank Tony Thomas for supplying me with the details of this paper during the Workshop, my actual talk featured an *Australian Journal of Physics* paper of his from 1971.

[2] "Metadata" refers to information not contained in a paper but rather about a paper, such as author, date, title, where it was published, etc. In other words the information that makes up a typical SPIRES record.

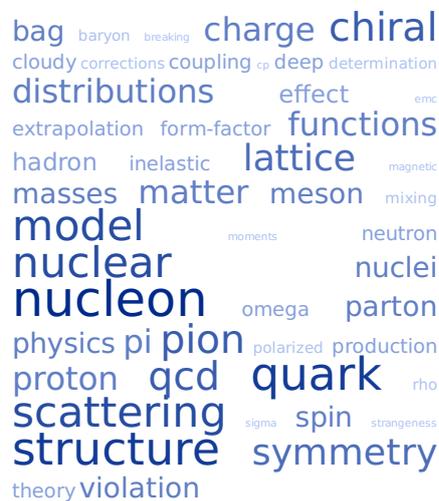**FIGURE 2.** The SPIRES record of Tony Thomas' first published paper.

FIGURE 3. Word cloud created from the title words of Tony Thomas' papers. This gives an easy-to-interpret representation of the importance of various topics in his work.

A BRIEF HISTORY OF SPIRES

A less brief history can be found elsewhere [1] but the following is, as far as I know, the first time the development of SPIRES has been discussed in parallel with the career of Tony Thomas.

In the late 1960s, as Tony entered undergrad at the newly built Flinders University, inaugural SLAC Director, Pief Panofsky charged the SLAC Library with cataloging the HEP preprints it received from around the world into the new computer database developed at Stanford University, SPIRES (the Stanford Physics Information REtrieval Sys-

tem[3]). SLAC would then send a weekly hardcopy mailing, listing all the new preprints, to universities and laboratories around the world. Concurrently the DESY Library was cataloging information on published papers in the HEPI (HEP Index) database.

The first half 1970s saw the Libraries of SLAC and DESY work together to create a single database, containing information on both preprints and published articles (with significant attention paid to matching one with the other). In addition to the preprint mailing, an "anti-preprint" list would also be sent, telling subscribers where various preprints from earlier mailings had been published. Around the time of this innovation, Tony was awarded his Ph.D. and left Australia for Canada, where he took up a postdoctoral position at the University of British Columbia. He, like the SPIRES database, now resided on the west coast of the North American continent.

A decade later, Tony returned to South Australia to take up a position as Professor of Physics at the University of Adelaide. It was around this time that SLAC introduced "Remote SPIRES" which enabled users around the world to communicate directly with the SPIRES database through email or bitnet login, enabling them to take advantage of the fact that SPIRES was a database that they could search in many different ways, rather than just experience it in an extremely limited fashion through the list of new preprints (which was now being sent by email as well as paper).

August of 1991 saw the introduction by Los Alamos string theorist, Paul Ginsparg, of the eprint[4] service xxx.lanl.gov. Physicists could place their TeXfiles there (first by email then later by ftp as well) and others could download them, providing instant access to the newest work. The SLAC Library helped by providing its preprint listserv to be used to notify the community of the day's eprints. In the same year, SLAC scientist Paul Kunz saw a demonstration of CERN engineer Tim Berners-Lee's newly invented World Wide Web. Upon return to SLAC, he told the SPIRES database manager, Louise Addis, that SPIRES would be an ideal candidate for the WWW and in December 1991 SPIRES became the first database on the Web and the first website outside of a handful of places in Europe. Berners-Lee referred to SPIRES as the "killer app" for the WWW [3].

The Los Alamos eprint server had initially been set up for formal theory papers but gradually began to include other fields, often with the help of servers at other institutions. In October of 1992 it was accepting papers in nuclear theory and the first Tony Thomas eprint appeared in January of 1993 [4]. In July of 1997 Adelaide launched a mirror site.

By the mid 2000s, as Tony Thomas had moved to Jefferson Lab to become its Chief Scientist, the SPIRES collaboration was a mature concern, with the workload shared by SLAC, DESY and Fermilab. However, the database's aging technological infrastructure made it difficult for developers to work with it. For users it was often painfully slow and far behind the state of the art for websites (though perhaps not so bad by the standards of academic websites [5]). Something had to be done.

Emerging from a 2007 meeting[5] of information professionals from CERN, DESY, Fermilab and SLAC, along with represetatives of the major HEP publishers and other

---

[3] As this system later came to be used to as a general database system for the university the acronym was redefined to stand for the Stanford Public Information REtrieval System.

[4] A neologism based on "electronic pre-print."

[5] http://indico.cern.ch/conferenceDisplay.py?confId=11611

HEP information resources (PDG, NASA-ADS, Durham, arXiv, etc.) a project was formed to move SPIRES to a more modern platform. CERN had created an open-source information management system, Invenio.[6] Unlike SPIRES, Invenio had a team of developers, was written in a modern computer language (Python) and enabled users to do lightning-fast searches over the web. An extensive survey of the HEP community to ascertain its information needs was carried out by a newly formed collaboration comprising the libraries of CERN, DESY, Fermilab and SLAC [6]. By 2009, as Tony Thomas was returning to Australia to resume his appointment at Adelaide, the collaboration was working on INSPIRE (INvenio + SPIREs), a system that would combine the content and functionality of SPIRES with the support and modernity of Invenio.

INSPIRE

In early 2010 a test version of INSPIRE was released, `http://inspirebeta.net`, just in time for this workshop celebrating Tony's 60th birthday. The user will notice several new features taken from the Invenio, such as

- the "detailed record" of a paper, which provides information such as co-cited papers and the citation history of the paper.
- the ability to store associated material such as the data behind plots or Mathematica notebooks.
- author pages listing co-authors, affilations, keywords and a citation summary.
- Google[TM]-type searching, e.g. `thomas starting chiral`.

Attributing papers in INSPIRE

It has always been difficult in SPIRES for many people to uniquely and comprehensive specify all their papers because of ambiguity in author names. This can lead to people employing long and unstable search strings, e.g.

`find a smith, j and d > 1988 not aff SLAC not a smith, t j`

It would be a great service to authors and users if we could uniquely associate a person with all of his or her papers, without users having to resort to such convoluted searches. To this end we have taken several steps. We will be introducing "INSPIRE ID numbers" in the HEPNAMES database, so that we can uniquely identify a person as both an author in HEP and a person in HEPNAMES and start associating papers to this single individual. This will also form the backbone of a system we'll soon deploy where authors can help us resolve ambiguities and "claim" their papers.

The INSPIRE collaboration is working with the ORCID Initiative[7] to ensure that getting your information right in INSPIRE means getting it right with publishers and other bibliographic entities and vice versa. This will represent significant progress,

---

[6] Latin for "I find." `http://cdsware.cern.ch/invenio`

[7] `http://orcid.org/media/pdf/ORCID_Announcement.pdf`

authors will only have to worry about getting their information right in one place and all associations of authors and papers will flow seamlessly through the chain: institutes, preprints, INSPIRE, publised articles and other databases.

The first steps have already begun: we are studying how to exchange information more efficiently with arXiv and working with various large HEP experimental collaborations to ensure we correctly attribute papers to their (thousands of) authors. As an example, we are helping collaborations include an `authors.xml` file with their arXiv submission, listing the INSPIRE numbers of all the authors on the paper so that nothing will be lost in the process. For example,

```
<authorOrganization>
  <author>
    <name>Thomas, A.W.</name>
    <authorID>INSPIRE-00131290</authorID>
  </author>
  <organization>adelaide.edu.au</organization>
<authorOrganization>
```

The near future

INSPIRE is due for official release in April 2010. Shortly after this we plan to introduce peronal accounts on INSPIRE allowing users to:

- Store favourite papers and set email alerts on pre-determined searches.
- Tag records with keywords and other information and correct mistakes more easily.
- Upload full-text of older papers (not suitable for arXiv) and deposit associated materials with existing papers.
- Display plots from papers and search figure captions.

ACKNOWLEDGMENTS

REFERENCES

1. H. B. O'Connell, *HEP Lib. Web.* **6**, 3 (2002) [arXiv:physics/0007040].
2. J. S. Schwinger, *Phys. Rev.* **82**, 664 (1951);
3. J. M. Deken, arXiv:physics/0208059.
4. W. Melnitchouk and A. W. Thomas, *Phys. Rev. D* **47**, 3783 (1993) [arXiv:nucl-th/9301016].
5. P. A. Kreitz and T. C. Brooks, *Sci. Tech. Libraries* **24**, 153 (2003) [arXiv:physics/0309027].
6. A. Gentil-Beccot *et al.*, *J. Am. Soc. Inf. Sci.* **60** (2009) 150 [arXiv:0804.2701 [cs.DL]].