Journal of Discrete Algorithms Special StringMasters Issue (2015)

Bill Smyth

PII: S1570-8667(15)00003-9
DOI: 10.1016/j.jda.2015.01.002
Reference: JDA 599

To appear in: *Journal of Discrete Algorithms*

Received date: 29 December 2014
Accepted date: 9 January 2015

Please cite this article in press as: B. Smyth, Journal of Discrete Algorithms Special StringMasters Issue (2015), *Journal of Discrete Algorithms* (2015), http://dx.doi.org/10.1016/j.jda.2015.01.002

Perspectives
Journal of Discrete Algorithms Special StringMasters Issue (2015)

After four previous journal issues of StringMasters[1] papers [1, 2, 3, 4], this fifth volume perhaps provides a felicitous moment to look back, to reflect on eight years of StringMasters events, more generally on a half-century of stringology/combinatorics-on-words in the age of the digital computer.

In July 2007 the eminent and (relatively) venerable Combinatorial Pattern Matching (CPM) symposium was to be held at the University of Western Ontario. UWO is about 90 minutes' drive from my university, McMaster, and for reasons now lost in the mists of time, it seemed to be a good idea to hold an informal meeting *chez nous* in the week prior to CPM. The first two paragraphs of the invitation sent out to several hundred researchers capture the spirit of those halcyon days:

> "Situated next to the sprawling Royal Botanical Gardens (RBG) and the picturesque town of Dundas, Ontario, McMaster University in the summertime could, admittedly only with a large leap of the imagination, be compared to Andalusia: flowers everywhere, ravines and small streams, tranquillity on an underpopulated student-deprived campus, profs transmogrified from pedants into peasants.
>
> "In keeping with this relaxed atmosphere, the Algorithms Research Group in the Department of Computing & Software (C&S) at McMaster proposes bringing together a selection of 'StringMasters' — *Die Meisterstringer* in German translation — and their students to work on string processing problems of current interest, especially those that relate to applications in data compression, data mining, computational biology, and

[1] It should be recorded for posterity that the title "StringMasters" was the inspiration of Maxime Crochemore.

musicology. Our discussions might well produce a ***string quartette*** — papers relevant to each of these four areas!"

It was surprising enough to me that our initiative attracted about 20 researchers and/or their students, flotsam from four continents; but since that time I have been particularly astonished at the continued and growing popularity of what I had thought was a one-off event — from a single repeat at McMaster in 2008 to four at diverse locations in 2014, with venues in between all over the world: Australia, South Africa, several universities in Europe[2]. At the same time, the five special issues of refereed research articles, four published annually since 2012, speak to the quality of the work done. Perhaps most surprising of all, the idea has spread to at least one other discipline: since 2010 there have been 11 GraphMasters workshops[3], devoted to open problems in graph theory.

It seems this new workshop paradigm fulfills a basic research need: rather than going to a meeting with a finished product, researchers instead turn up with ideas and conjectures, share them with colleagues, work in partnership to refine them into theorems and algorithms. Does the success of StringMasters say something about the collegial nature of our work, does it derive from a shared respect for some mathematical or computational Holy Grail that we all search for?

The success of the StringMasters phenomenon surely also has something to do with the maturation of our discipline, one that still does not have an agreed-upon title: "combinatorics on words", "stringology", and "algorithms on strings" overlap hugely, and there are close links to formal language theory, finite automata theory, and various application areas — text compression, bioinformatics, natural language processing, among others. Uniquely named or not, the study and influence of *stringology* (let us call it) has certainly exploded in the quarter-century that I have been working in the field, and it had been steadily expanding, driven by numerous applications, for another quarter-century before that (remember [5], celebrating the "myriad", now largely outdated, uses of suffix trees?).

Outdated, no doubt, but only by improved methods and deeper insights that it took many years for an entire new generation of stringologists to discover. The suffix tree has been largely supplanted by the suffix array, first proposed in 1990, but not used extensively in applications until 15–20 years later, when efficient methods for its computation and use were gradually made clear by the accumulated efforts of hundreds of researchers. And this is but one of numerous areas of interest, both applied and theoretical, in

[2]For a complete list see `http://www.stringmasters.org/main_previous.php`.
[3]See `http://www.graphtheorygroup.com/graphmasters/`.

which marvellous discoveries have been made over the last half-century or so. Here are some of my own favorites:

- the amazing progress made in the specification of compressed data structures and their efficient use in string algorithms;
- the discovery of a remarkable dozen-line algorithm to decompose a string into an ordered sequence of Lyndon words;
- the very recent and very simple proof, using Lyndon words, that the number of maximal periodicities in a string cannot exceed the string's length, a problem intensively studied by dozens of researchers over the last 15 years;
- numerous results on the properties and processing of "indeterminate" (or "generalized" or "degenerate") strings — those whose entries consist of subsets, rather than elements, of a given alphabet;
- more precise understanding of the combinatorial properties of strings in the neighbourhood of a "double square", and potential application to the design of greatly improved algorithms to compute repetitions;
- after hundreds, if not thousands, of research papers, continued progress on the pattern matching problem: find all occurrences of a given pattern in a given string.

Another stringologist with different interests would come up with a list disjoint from mine, but no less impressive.

Some measure of the growth and reorientation of stringology may be inferred by considering conferences in the area. Before 1990 I know of only one: the International Conference on Automata & Formal Languages (AFL), established in 1980. The next to appear was the "venerable" CPM in 1990, and here is a list of acronyms of 16 conferences that have been established since, together with their founding dates:

> DCC (1991), SPIRE (1994), PSC (1996), CIAA(1996), WORDS (1997), DLT (1997), WABI (2001), LSD/LAW (2002), ECCB (2002), FSMNLP (2002), WCTA (2006), PRIB (2006), Stringology Research Workshop (2006), LATA (2007), StringMasters (2007), SISAP (2008).

Perhaps I have included a meeting or two that I shouldn't, perhaps I have missed a couple. Overall, there can be little doubt that by this measure stringology is expanding at a gallop.

Another important indicator of the maturity of a research field is surely the number of books published. The first text on graph theory, published in German by Dénes König in 1936, was famously influential in rescuing graph theory from the "slums of topology". Today there must be hundreds of textbooks available on aspects of graph theory and graph algorithms. A similar role has doubtless been played in stringology by [6], a collaborative foundational effort by many authors to express the important ideas and results in our field. This text was the only work on stringology available in the 1980s; in my bookcase I count three that were published in the 1990s, and since 2001 there have been at least a dozen, including two further substantial volumes in the Lothaire series [7, 8]. Indeed, it should not be surprising that graph theory and stringology have similar developmental histories: the mathematical basis of each is combinatorial, expressed in deceptively simple models, and both have been greatly popularized by computer-based applications: for graph theory in "operations research" initially made urgent (on both sides) by the requirements of World War II, for stringology in bioinformatics spurred by Crick & Watson's 1953 discovery of DNA structure, the four-letter "code of life".

In this context, it should not go unremarked that very recently, at end 2014, "String Algorithmics & Applications" has been created as Working Group 1.10 within Technical Committee 1 of the International Federation for Information Processing[4]. Founded in 1960, IFIP represents IT societies from 56 countries around the world, with a total membership of over half a million. At a more modest level, Murdoch University in Western Australia has recently established a Centre for Combinatorics on Words & Applications (CCWA)[5], a first step in linking together stringologists with a common (scientific) language and shared interests.

When I began working in stringology in 1989, there were perhaps a couple of hundred researchers in the field; now there must be several thousand, plus no doubt an equivalent number of bioinformaticians to whom suffix trees and suffix arrays are as familiar as introns and exons. The Journal of Graph Theory[6] was founded in 1977 when the number of active graph theorists was certainly no more than the number of stringologists today; but from the beginning it was a successful innovation, primarily because it related, as does stringology, both to numerous applications and to fundamental combinatorial problems. Perhaps the time has come for a "Journal for Stringology & Applications" to be established, to provide an intellectual home, and inspiration, for the best research our field has to offer. Indeed, this is one of the stated purposes of the CCWA, mentioned above.

[4]See `http://www.ifip.org/index.php`.
[5]See `https://ccwaperth.wordpress.com/`.
[6]See `http://www3.interscience.wiley.com/journal/35334/home`.

Bill Smyth
Department of Computing & Software
McMaster University, Hamilton, Canada

Department of Informatics
King's College London, UK

School of Engineering & Information Technology
Murdoch University, Perth, Western Australia

References

[1] Ryszard Janicki, Simon J. Puglisi & M. Sohel Rahman (eds.), Special Issue on Stringology, *Fundamenta Informaticae 97–3* (2009) 52 pp.

[2] Jacqueline W. Daykin, Bill Smyth & Lynette van Zijl (eds.), StringMasters 2009 & 2010 Special Issue, *J. Discrete Algs. 10* (2012) 34 pp.

[3] Maxime Crochemore & Jacqueline W. Daykin (eds.), StringMasters 2011 Special Issue, *J. Discrete Algs. 20* (2013) 70 pp.

[4] Maxime Crochemore, Jacqueline W. Daykin & Zsuzsanna Lipták (eds.), StringMasters 2012 & 2013 Special Issue (Volume 1), *J. Discrete Algs. 28* (2014) 84 pp.

[5] Alberto Apostolico, **The myriad virtues of suffix trees**, *Combinatorial Algorithms on Words* (NATO ASI Series F12), Springer-Verlag (1985) 85–96.

[6] M. Lothaire, *Combinatorics on Words*, Encyclopædia of Mathematics, vol. 17, Addison-Wesley (1983) 260 pp.

[7] M. Lothaire, *Algebraic Combinatorics on Words*, Cambridge University Press (2002) 504 pp.

[8] M. Lothaire, *Applied Combinatorics on Words*, Cambridge University Press (2005) 610 pp.