Invited Editorial

Manually structured digital abstracts: A scaffold for automatic text mining

In the past, we have advocated the adoption of the structured digital abstract to bring scientific publishing into the database age [1,2]. An increasing number of projects are bringing us toward the reality of machine-readable as well as human-readable access and integration to large published data sets, and as such we will take a moment to revisit our proposal, reflect and address several of the concerns that have arisen since our articles first appeared last year.

In brief, we envision the structured digital abstract as applying text-mining software (with curatorial supervision, where necessary) to accepted journal articles at the pre-print stage. This process will be carried out by the journal as a normal step in the publication process, and authors will then use this output as a starting point to shape the digital abstract. The final structured abstract – a machine-readable snapshot of the soon-to-be-published data – is subject to editorial approval (and eventually peer review), to assure proper and accurate classification and tagging of data. Ultimately, we envision the structured abstract becoming an integral step and accepted in the publication process, much in the way that scientists must now spend time formatting an article for a specific journal. With respect to text-mining techniques, these manually-verified final structured abstracts will be invaluable in providing gold-standard data sets for training and refining text-mining algorithms.

It has been suggested that asking authors to vet the structured digital abstract imposes an additional burden on the editorial process [3]. It may well be that the SDA requires the efforts of authors, editors and the additional input of a curator versed in the particular content descriptors for a given species or subject of research. But compared with the backward system of curator-only retroactive text-mining, generation of the abstract at the time of publication will produce a more accurate and useful computer-ready companion to the paper. Moreover, text mining is far more effective when armed with a robust translation table generated with author input (i.e. a list of gene names) rather than by *post hoc* text-mining approaches. The structured abstract will provide valuable 'context' to mining algorithms by presenting clearly the main points of each article (as defined by authors and editors), so additional facts gleaned can be correctly categorized as either supporting or detracting from the main points.

Another potential pitfall is the fragmentary coverage of existing terminology systems. Gaps in terminology pose problems for a systematized markup scheme. We believe the solution to this – the best way to expand such systems – is to let authors contribute. Curators provide indispensable expertise in categorizing and labeling data, but it is unrealistic to expect curators to maintain personal familiarity with the vast array of facts and concepts in biology. The way to encompass all needed terminologies is to involve the entire research community. Authors are heavily invested in their papers, with a strong interest in making sure data are represented correctly. With author contribution, existing gaps in ontological coverage should shrink rather quickly. The structured digital abstract system, then, may well be challenged by such gaps in terminology – but it is also the best way to patch them.

A potential issue that Hahn et al. [3] point to is the tendency of authors to be subjective, perhaps overly positive. Peer review of the nascent digital abstract should combat any puffery. Current classification approaches are neither mandatory nor peer-reviewed, nor implemented automatically at publication. Peer review is essential to preserving scientific integrity, and it is for this reason that we have always advocated hatching digital abstracts under its purview.

Hahn et al. [3] conclude that an alternative solution is automatic text mining. Text mining is important – indeed, it is the bedrock of the structured digital abstract initiative. But we envision journals themselves spearheading this initiative, invoking the latest text-mining software at the pre-print stage and subjecting the digital abstract to author- and peer-review. This strategy can be implemented immediately, even with imperfect text-mining software, as opposed to waiting until sufficient progress has been made toward a fully automated solution.

The FEBS Letters experiment toward integrating human-readable output with large-scale protein data sets is encouraging. We look forward to this exciting (and imminent) reality, where smooth data integration and machine-readable abstracts bring the vast and growing corpus of scientific literature within reach of our most powerful data mining and access tools.

References

[1] Gerstein, M., Seringhaus, M. and Fields, S. (2007) Structured digital abstract makes text mining easy. Nature 447, 142.

[2] Seringhaus, M.R. and Gerstein, M.B. (2007) Publishing perishing? Towards tomorrow's information architecture. BMC Bioinform. 8, 17.

[3] Hahn, U., Wermter, J., Blasczyk, R. and Horn, P.A. (2007) Text mining: powering the database revolution. Nature 448, 130.

Michael Seringhaus, Mark Gerstein
Yale University, Department of Molecular Biophysics and Biochemistry,
New Haven,
CT 06520,
USA
E-mail address: Mark.Gerstein@Yale.Edu

0014-5793/$34.00 © 2008 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2008.02.073