EXECUTIVE SUMMARY
How can life science researchers stay on top of the constantly growing body of Medline-indexed articles that are potentially relevant to their work? Reading through the more than one million such articles published annually clearly is not an option. That leaves two primary strategies for sifting through the burgeoning literature and extracting meaningful information: manual curation or automated curation.
For years, manual curation of scientific publications has been the gold standard, with technology-based solutions ranking far behind in terms of accuracy and completeness. Today, that is no longer the case. Versatile, well-designed and well-tested applications combined with significantly enhanced computational power are elevating automated curation to a more equivalent position; proprietary text-mining technologies now rival manual curation as a means of ensuring that researchers are not missing out on valuable information. Which solution works best for a given researcher, laboratory or organization varies. Here are some key factors to consider when you are deciding on automated vs. manual text mining.

**Scope and Volume**
The information needed to interpret experimental results, describe a cellular pathway or identify complex interactions in regulatory networks often is scattered throughout hundreds of articles and publications—more than a single researcher can review. That can leave you with a gnawing feeling of “What have I missed?”

The solution is to cast a wide net, reviewing as many articles from as many journals as possible across a specific field. But that’s simply not practical in most cases. One option is to read only the abstracts of papers. While these are considerably shorter, they do not contain all the important information found in the full text. Multiple studies comparing the full text and abstract from the same paper concluded that less than half of the key facts from the body of a paper are present in the abstract.

Another option is to rely on PhD researchers trained as manual curators. However, those experts can read and annotate only about 20-25 papers a day—fine, perhaps, for a highly targeted query to a small number of journals in a select area, but not adequate for comprehensive coverage of a topic.

Bias also is a factor when people have to wade through an extraordinary amount of data. Manual curation can introduce bias by limiting journals and articles due to resource restrictions and assumptions about journal value. Automated systems scan all Medline abstracts (without bias) and millions of full-text articles (the limit here mainly is due to legal issues and licensing fees).

The other “bias” argument is that manual curators cull only the most appropriate articles from high-profile journals in any given field. Yet, these days, critical information regarding particular pathways or relationships can turn up just about anywhere. The ability to rapidly process millions of Medline abstracts and full-text articles from quality journals substantially increases the odds of capturing relevant data.

**Accuracy**
Some would argue that quality is more important than quantity—and that manual curation ensures accuracy. Yet research has shown that manual curation is not perfect. Overall, expert curators are about 90% accurate (as measured by inter-curator agreement on annotation) for specific tasks, and inter-annotator agreement ranges from 77% to a low of 31%. In the past five to seven years, the accuracy of specialized automated text-mining systems has improved dramatically. In-house research at Elsevier reveals our automated system’s accuracy is about 82-85% overall. Moreover, automated systems can be adapted to include new terms and concepts in biology rapidly, simply by adding new ontologies. In addition, automated systems are exceptionally consistent in their annotation from paper to paper and journal to journal, unlike human curators who show some natural variation over time.

**Speed**
In many situations, speed is of the essence—e.g., if a researcher needs specific information to meet a grant proposal deadline, or to reduce the time required to get a new drug to market. In this realm, there’s no contest: a trained...
In biological research, identifying relationships between entities—e.g., protein-protein or drug-protein interactions—is at the heart of pathway analysis. To do this effectively, automated curation would have to be able to mimic the human ability to infer connections from text—and indeed it can. Elsevier’s natural language text-mining system can identify meaningful relationships through a combination of specialized ontologies and linguistics rules, much the same way humans identify relationships through reading.

Although abstracts are short, and many can be read quickly, they don’t contain all the key facts from the full-text publication. Typically less than half of cited terms in a paper are mentioned in the abstract. Because this automated system scans full-text articles as well as abstracts, it can identify many more relevant relationships than could be found by scanning abstracts alone.

Some researchers feel comfortable relying on a curator’s expertise to identify important information in the literature. Others want to at least be able to review the results and use their own expertise to decide whether a particular finding or relationship is relevant or credible. Automated systems such as Elsevier’s do something most systems based on manual curation don’t: they show the sentence in the abstract or paper used to identify each relationship, so the researcher can personally review them and decide whether or not to include or exclude any reference. So rather than relying on someone else’s judgment, the final decision about what information to believe for their research rests with each user.

If you’re trying to decide between systems based on manual and automated curation, ask yourself: Would your project benefit from information obtained from a wide swath of journals or just a chosen few? How long are you willing to wait to get access to information? Does identifying a greater number of relevant relationships between entities give you more confidence in your data? Are you comfortable letting others decide what research is most relevant to your work, or do you want to review the information and make that decision yourself?

---

Interested in learning how automatic curation can improve your target-based research?

Discover more at
www.elsevier.com/pathway-studio

ASIA AND AUSTRALIA
Tel: + 65 6349 0222
Email: sginfo@elsevier.com

JAPAN
Tel: + 81 3 5561 5034
Email: jpinfo@elsevier.com

KOREA AND TAIWAN
Tel: +82 2 6714 3000
Email: krinfo.corp@elsevier.com

EUROPE, MIDDLE EAST AND AFRICA
Tel: +31 20 485 3767
Email: nlinfo@elsevier.com

NORTH AMERICA, CENTRAL AMERICA AND CANADA
Tel: +1 888 615 4500
Email: usinfo@elsevier.com

SOUTH AMERICA
Tel: +55 21 3970 9300
Email: brinfo@elsevier.com

Elsevier is a registered trademark of Elsevier BV Pathway Studio is a registered trademark of Elsevier Inc.