Whitepaper

Semantic Enrichment and the Information Manager

Turning Content into Insight

Michael Iarrobino, CCC
Lee Harland, SciBite

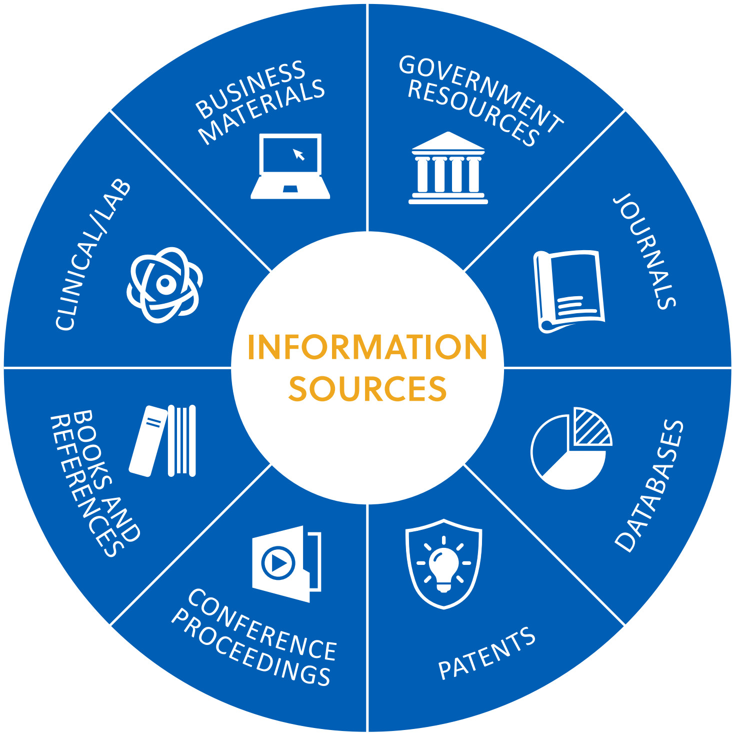
It is more resource intensive to bring a drug to market than ever before: Pharmaceutical companies spend an average of 10+ years and $2.6 billion on each successful outcome.[1] Fewer than 1% of potential drugs are ever made available for sale.[2]

Professionals across the drug development pipeline know well the daily struggle of staying current with a multiplying volume of multitudinous types of information: peer-reviewed scientific research, patent filings, clinical trials data, news and competitive briefs, conference abstracts and posters, and more.

At the same time, the growth in volume of published research is accelerating: More scientific research published between 2010 and 2014 was indexed in MEDLINE than all research published before 1970.[3]

Professionals across the drug development pipeline know well the daily struggle of staying current with a multiplying volume of multitudinous types of information: peer-reviewed scientific research, patent filings, clinical trials data, news and competitive briefs, conference abstracts and posters, and more.

This paper explores these challenges, and how a core concept—*semantic enrichment*—can be a foundation to address these problems and open new opportunities to quickly extract insights from information.



The information management evolution

Life sciences and pharmaceutical companies are increasingly turning to informatics teams to lead initiatives in artificial intelligence, cognitive search, and machine learning aimed at exploiting vast amounts of internal and external information.

Although initiatives like these show great promise, the reality is often less promising. Organizations struggle with interoperability across different data streams. The US healthcare industry has captured just **10–20%** of the potential value from big data analytics, according to McKinsey Analytics.[4]

At the same time, information management is evolving. Information managers must balance the needs of multiple internal constituencies to support information discovery and manage subscription usage and budgets. Enhanced search, personalization, and collaborative workflows should all be top priorities.

The challenge is that **more than 90%** of organizational data is unstructured,[5] as is nearly all the peer-reviewed scientific research. Although it's true that articles from journal subscriptions have a semi-structure of metadata and discrete article sections, actionable intelligence lies trapped in the unstructured research narratives of the full text.

As information management professionals consider approaches to make these insights easily discoverable and accessible, and to extract increased value from journal subscriptions and other unstructured content, their purposes may converge with those of their informatics colleagues. Although informatics teams may focus more on applying machine analysis methods, such as text mining, to content, the shared goal is to deliver actionable intelligence across the organization.

To advance such efforts, informatics teams contribute expertise in data transformation and interoperability. Information management brings a strong awareness of how content consumers in the organization prefer to discover and access content, depth of knowledge of the sources of externally published scientific research, and expertise regarding the efficiency, effectiveness, and comprehensiveness of information discovery.

Enhanced search, personalization, and collaborative workflows should all be top priorities.

How semantic enrichment can help

Semantic enrichment is the enhancement of content with information about its meaning, thereby adding structure to unstructured information.[6] While unstructured content must be synthesized from scratch each time it is consumed, semantically enriched content has been annotated with its meaning, enabling users to move quickly to more intelligence-rich information activities.

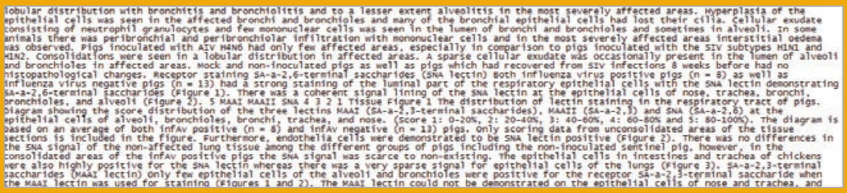
Semantic enrichment can annotate unstructured text with information that links related concepts together and makes clear where these concepts sit in a hierarchical family. Think of the many ways to refer to cancers or neoplasms, and the types that are subordinate to the overall cancer concept. This enables a user looking for information about cancer as a concept to gather the many different references of it as a class, achieving a recall impossible with information in the unstructured text alone.

Semantic enrichment can disambiguate unstructured text. For example, consider the importance of being able to differentiate AIDS (Acquired Immunodeficiency Syndrome) from hearing aids. A user searching for aids using keywords will likely retrieve examples of both, while a user who can distinguish between the concepts achieves greater precision.

Semantic enrichment is a key enabler of the various strategic initiatives undertaken by informatics and information management professionals. So how does it work?

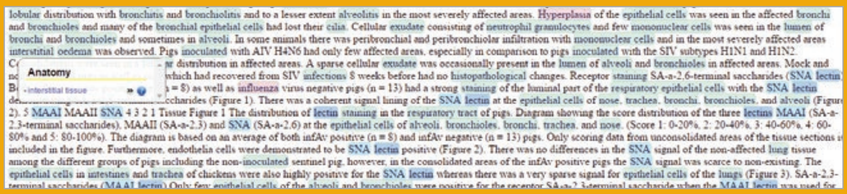
> Semantic enrichment is a key enabler of the various strategic initiatives undertaken by informatics and information management professionals.

Before Semantic Enrichment

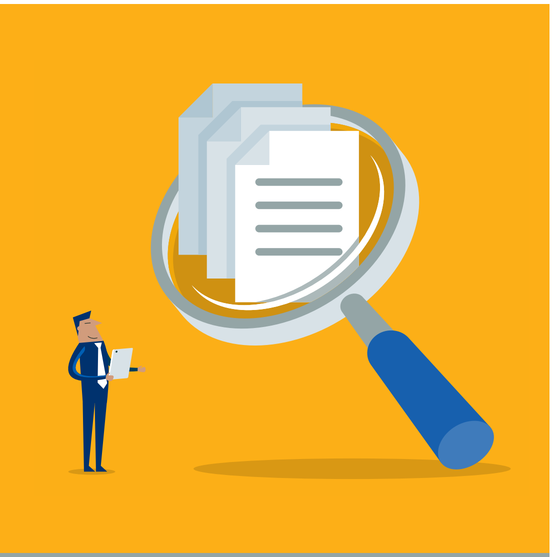


lobular distribution with bronchitis and bronchiolitis and to a lesser extent alveolitis in the most severely affected areas. Hyperplasia of the epithelial cells was seen in the affected bronchi and bronchioles and many of the bronchial epithelial cells had lost their cilia. Cellular exudate consisting of neutrophil granulocytes and few mononuclear cells was seen in the lumen of bronchi and bronchioles and sometimes in alveoli. In some animals there was peribronchial and peribronchiolar infiltration with mononuclear cells and in the most severely affected areas interstitial oedema was observed. Pigs inoculated with AIV H4N6 had only few affected areas, especially in comparison to pigs inoculated with the SIV subtypes H1N1 and H1N2. Consolidations were seen in a lobular distribution in affected areas. A sparse cellular exudate was occasionally present in the lumen of alveoli and bronchioles in affected areas. Mock and non-inoculated pigs as well as pigs which had recovered from SIV infections 8 weeks before had no histopathological changes. Receptor staining SA-a-2,6-terminal saccharides (SNA lectin) Both influenza virus positive pigs (n = 8) as well as influenza virus negative pigs (n = 13) had a strong staining of the luminal part of the respiratory epithelial cells with the SNA lectin demonstrating SA-a-2,6-terminal saccharides (Figure 1). There was a coherent signal lining of the SNA lectin at the epithelial cells of nose, trachea, bronchi, bronchioles, and alveoli (Figure 2). 5 MAAI MAAII SNA 4 3 2 1 Tissue Figure 1 The distribution of lectin staining in the respiratory tract of pigs. Diagram showing the score distribution of the three lectins MAAI (SA-a-2,3-terminal saccharides), MAAII (SA-a-2,3) and SNA (SA-a-2,6) at the epithelial cells of alveoli, bronchioles, bronchi, trachea, and nose. (Score 1: 0-20%, 2: 20-40%, 3: 40-60%, 4: 60-80% and 5: 80-100%). The diagram is based on an average of both infAv positive (n = 8) and infAv negative (n = 13) pigs. Only scoring data from unconsolidated areas of the tissue sections is included in the figure. Furthermore, endothelia cells were demonstrated to be SNA lectin positive (Figure 2). There was no differences in the SNA signal of the non-affected lung tissue among the different groups of pigs including the non-inoculated sentinel pig, however, in the consolidated areas of the infAv positive pigs the SNA signal was scarce to non-existing. The epithelial cells in intestines and trachea of chickens were also highly positive for the SNA lectin whereas there was a very sparse signal for epithelial cells of the lungs (Figure 3). SA-a-2,3-terminal saccharides (MAAI lectin) only few epithelial cells of the alveoli and bronchioles were positive for the receptor SA-a-2,3-terminal saccharide when the MAAI lectin was used for staining (Figures 1 and 2). The MAAI lectin could not be demonstrated on the epithelial cells of nose and trachea, and

After Semantic Enrichment



lobular distribution with bronchitis and bronchiolitis and to a lesser extent alveolitis in the most severely affected areas. Hyperplasia of the epithelial cells was seen in the affected bronchi and bronchioles and many of the bronchial epithelial cells had lost their cilia. Cellular exudate consisting of neutrophil granulocytes and few mononuclear cells was seen in the lumen of bronchi and bronchioles and sometimes in alveoli. In some animals there was peribronchial and peribronchiolar infiltration with mononuclear cells and in the most severely affected areas interstitial oedema was observed. Pigs inoculated with AIV H4N6 had only few affected areas, especially in comparison to pigs inoculated with the SIV subtypes H1N1 and H1N2.
C ar distribution in affected areas. A sparse cellular exudate was occasionally present in the lumen of alveoli and bronchioles in affected areas. Mock and
no which had recovered from SIV infections 8 weeks before had no histopathological changes. Receptor staining SA-a-2,6-terminal saccharides (SNA lectin)
B n = 8) as well as influenza virus negative pigs (n = 13) had a strong staining of the luminal part of the respiratory epithelial cells with the SNA lectin
de ccharides (Figure 1). There was a coherent signal lining of the SNA lectin at the epithelial cells of nose, trachea, bronchi, bronchioles, and alveoli (Figure 2). 5 MAAI MAAII SNA 4 3 2 1 Tissue Figure 1 The distribution of lectin staining in the respiratory tract of pigs. Diagram showing the score distribution of the three lectins MAAI (SA-a-2,3-terminal saccharides), MAAII (SA-a-2,3) and SNA (SA-a-2,6) at the epithelial cells of alveoli, bronchioles, bronchi, trachea, and nose. (Score 1: 0-20%, 2: 20-40%, 3: 40-60%, 4: 60-80% and 5: 80-100%). The diagram is based on an average of both infAv positive (n = 8) and infAv negative (n = 13) pigs. Only scoring data from unconsolidated areas of the tissue sections is included in the figure. Furthermore, endothelia cells were demonstrated to be SNA lectin positive (Figure 2). There was no differences in the SNA signal of the non-affected lung tissue among the different groups of pigs including the non-inoculated sentinel pig, however, in the consolidated areas of the infAv positive pigs the SNA signal was scarce to non-existing. The epithelial cells in intestines and trachea of chickens were also highly positive for the SNA lectin whereas there was a very sparse signal for epithelial cells of the lungs (Figure 3). SA-a-2,3-terminal saccharides (MAAI lectin) Only few epithelial cells of the alveoli and bronchioles were positive for the receptor SA-a-2,3-terminal saccharide when the MAAI lectin was used for



Semantic enrichment opens possibilities to informatics and information management that couldn't have been considered before.

Semantic enrichment requires these raw materials

Content. Semantic enrichment can be conducted on disparate unstructured internal and external content, including journal article subscriptions.

Vocabulary. There are myriad publicly available vocabularies, including Medical Subject Headings (MeSH), RxNorm, MedDRA, SNOMED, and more; in the biomedical domain alone, more than 500 vocabularies may be applied, encompassing more than 7 million classes.[7] An organization may also extend these or create its own vocabularies to tailor the enrichment to its user needs.

Rules. These are instructions for how to apply the vocabulary to unstructured content.

Semantic enrichment can also unify different sources of unstructured content with a single semantic layer. For example, an internal clinical report that has been annotated to apply MeSH to the various medical concepts discussed can be readily compared with the full text of a peer-reviewed journal article that has been annotated using the same vocabulary and process.

The semantically enriched content can then be used to feed a variety of downstream information consumers, from search applications, triplestores, and graph databases to content alerting systems and analytics and visualizations tools.

The benefits of semantic enrichment

Semantic enrichment opens possibilities to informatics and information management that couldn't have been considered before. Keyword search and manual curation could never have approached the task of examining the more than 25 million articles indexed in MEDLINE to extract candidate relationships between whole classes of concepts like genes and diseases, irrespective of phrasing, location, source, and format. But semantic enrichment can exploit content in this way.

Semantic enrichment also helps solve many already existing and intractable problems across different functional areas and teams, including:

- **Early phase research.** Semantically enriched content annotated with relevant biological, disease, protein, and gene concepts can be analyzed to determine potential relationships between these. The resulting relationship graph can suggest potential biomarkers and drug targets, and the findings linked to supporting source content for validation prior to wet lab.
- **Competitive intelligence.** Competitor patent filings, often intended to hinder discovery, can support improved recall through semantic enrichment that enhances text to annotate chemical substances. Non-patent literature (NPL) enriched using the same vocabularies can be explored alongside the patent literature to provide a full picture for patent landscaping or other competitive purposes.
- **Pharmacovigilance.** Scientific research semantically enriched to identify adverse events and pharmacological substances can suggest links between the two, increasing the efficiency and comprehensiveness of these vital monitoring workflows.
- **IDMP (identification of medicinal products) compliance.** IDMP initiatives directed by the Food and Drug Administration (FDA) and European Medicines Agency (EMA) aim to standardize how information can be expressed about pharmacological products. When semantically enriched, disparate internal and external content sources can be exploited to give a more robust view of product attributes.

Semantic enrichment and scientific literature: three takeaways for information managers

Although semantic enrichment is a complex process, it produces powerfully simple business results. For information managers, it has the ability to reduce friction in the discovery, access, consumption, and synthesis of published scientific literature. Here are three takeaways:

Semantic enrichment is a complex process, producing powerfully simple business results.

1 Driving discovery

Published journal content is typically enhanced with layers of metadata — abstract, publication type, date, keyword and topic categories, and many more — to enhance discoverability. Semantic enrichment provides a new approach that can go deeper than these traditional methods to expose the relevant concepts present in the full text of an article.

2 Measuring the value of content

Information managers continually analyze subscription and document delivery usage, adjusting their content sourcing as needed to meet evolving needs. Externally published literature that has been semantically enriched can be used and consumed in new ways, sometimes even without direct human involvement, and thus may not be properly represented in traditional models of article views, downloads, or purchases. Information managers may need new approaches and metrics to understand the value that published content delivers to their organization.

3 Enhancing the role of information management

Information managers have a unique perspective and expertise to contribute to semantic enrichment projects. Look for opportunities to partner with other groups, such as informatics, to contribute awareness of the ways the organization currently acquires and uses published content.

Michael Iarrobino

Copyright Clearance Center

Michael Iarrobino is Product Manager at Copyright Clearance Center (CCC), the leader in content workflow and rights licensing technology. He oversees CCC's RightFind® XML for Mining product, a workflow solution for text mining researchers using peer-reviewed scientific articles. He has previously managed products that solve problems in the marketing technology and content discovery spaces while at FreshAddress, Inc., and HCPro, Inc. He speaks at webinars and conferences on the topics of content discovery and data management.

Lee Harland

SciBite

Lee Harland founded SciBite to address a gap in robust, industry-centric text analytics solutions. He has extensive experience in life sciences with a PhD in genetics from Kings College London, followed by more than 15 years leading semantic web, data integration and text-mining efforts as applied to industrial life sciences. He has published many papers in these areas and serves as an advisor and collaborator to a number of initiatives such as Open PHACTS, BiomedBridges, the Experimental Factor Ontology/Biosamples and ChEMBL groups.

Resources

[1] https://www.scientificamerican.com/article/cost-to-develop-new-pharmaceutical-drug-now-exceeds-2-5b/ ; http://csdd.tufts.edu/news/complete_story/pr_tufts_csdd_2014_cost_study

[2] http://pubs.rsc.org/en/content/chapterhtml/2015/bk9781782621898-00001?isbn=978-1-78262-189-8

[3] https://www.nlm.nih.gov/bsd/medline_lang_distr.html

[4] http://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world

[5] https://www.slideshare.net/cscyphers/bd-101-slideshare

[6] Very good, more technical definition here: http://www.councilscienceeditors.org/wp-content/uploads/v37n2p40-44.pdf

[7] According to data indexed by National Center for Biomedical Ontology; http://bioportal.bioontology.org/s

Copyright Clearance Center (CCC)
A pioneer in voluntary collective licensing, Copyright Clearance Center (CCC) helps organizations integrate, access, and share information through licensing, content, software, and professional services. With expertise in copyright and information management, CCC and its subsidiary RightsDirect collaborate with stakeholders to design and deliver innovative information solutions that power decision-making by helping people integrate and navigate data sources and content assets.

CCC | RightsDirect

Learn more about our licensing, content, and data solutions:

U.S. organizations:
🌐 copyright.com/rightfind
✉ solutions@copyright.com

Outside U.S. organizations:
🌐 rightsdirect.com/rightfind
✉ solutions@rightsdirect.com