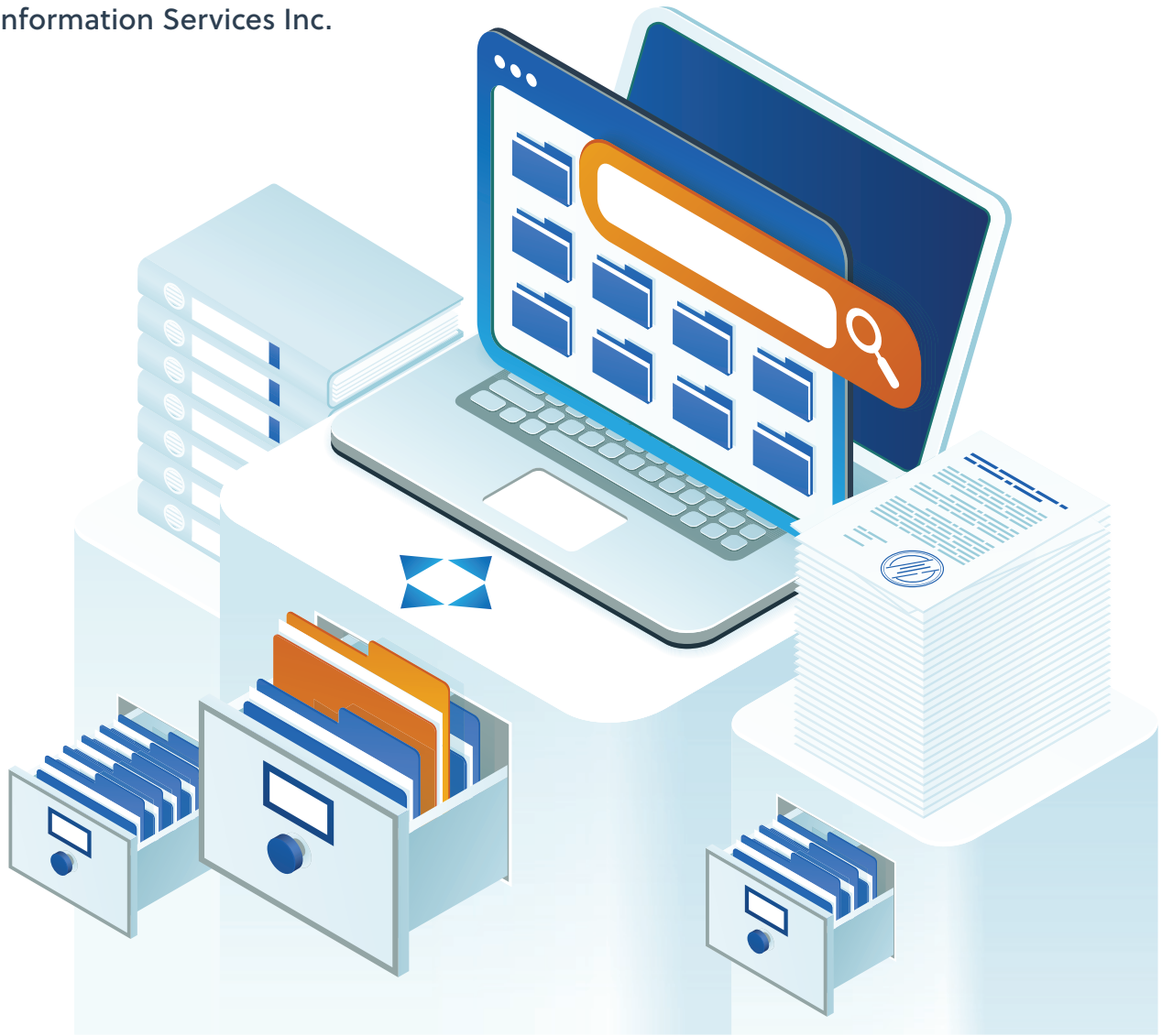
Copyright Clearance Center | RightsDirect

Whitepaper

Vocabularies, Text Mining and FAIR Data

The Strategic Role Information Managers Play

Mary Ellen Bates
Bates Information Services Inc.



One of the most important drivers in scientific knowledge discovery is the rapidly growing amount of data available. Peer-reviewed scientific research in more and more specialized fields, clinical trials data, conference proceedings and posters, business news and regulatory filings, patents, government datasets, and internal research and data all contain information valuable to researchers. However, much of this information is either unstructured or in data streams that do not permit easy interoperability. Adding to the challenge are internal data silos, often resulting from corporate acquisitions in which a group's data set is in a format or on a platform that the parent company does not use, and customized ontologies created or licensed by individual research teams. Organizations are beginning to recognize that simply having access to information does not ensure that the relevant information can be found or that relationships among items will be surfaced to accelerate new discoveries. Insight requires more than data and a search tool.

In this paper, we look at the points at which information managers can have the greatest impact serving as information catalysts — cleaning data, enhancing metadata, and managing the knowledge graphs that shape how an enterprise "sees" data. Then we look at the opportunities to advocate for the importance and impact of data cleansing and semantic enrichment to the enterprise's information flow.

Making data FAIR

Text and data mining initiatives, machine learning projects, and other artificial intelligence programs attempt to address this challenge by making metasearch possible, but these efforts require data and metadata that is:

- Consistent across internal and external data sources
- FAIR (Findable, Accessible, Interoperable, and Reusable)

The FAIR guiding principles, first articulated in a Scientific Data article in March 2016 are designed to make data both machine- and human-actionable:

- **Findability:** Metadata is assigned consistently and permanently, and maintained in a searchable source.
- **Accessibility:** Once identified, the data is accessible, given any necessary authentication and authorization; the metadata remains accessible even if the data is no longer available.
- **Interoperability:** The data and metadata can be integrated with other data, and can be used by applications for analysis and further processing; metadata can reference other metadata.
- **Reusability:** Metadata is applied to the data as thoroughly as possible, so that it can be reused in unanticipated ways and so that context and provenance is retained.

Information managers add FAIRness

Information managers bring a unique set of skills to the organization with their understanding of what information sources and tools are available, how information flows within the organization, and how various user groups acquire, use, and store information. They can have important consultative roles to play in every aspect of creating and maintaining FAIR data and workflow as outlined below.

- **Findability:** Information managers can identify existing internal ontologies and vocabularies, third-party ontologies, and semantic enrichment tools to consistently apply the right metadata to internal and external data. This significantly increases the ability of a user to find the necessary data, regardless of its source.
- **Accessibility:** Information managers are often tasked with identifying internal data sources and open source content, and licensing content from publishers and content providers. With their deep familiarity with the varied information needs and workflow processes of the organization, information managers bring a unique perspective to acquiring and licensing content.
- **Interoperability:** Information managers can work with research teams to bring semantic enrichment to internal data, licensed content, and data streams, and work to facilitate the sharing of information collections, APIs, and ontologies. This de-siloing of data and resources increases the ROI of the content and leverages enterprise investments.
- **Reusability:** Information managers have an enterprise-perspective on which groups could benefit from an information resource, and know how to leverage resources to make content acquisitions as cost-effective as possible. Their long familiarity with negotiating content licenses enables them to get better value for their content investments.

Information professionals have long been familiar with metadata—author, source, date, subject terms—enhancing the findability of published material such as journal articles or bibliographic citations. But this type of metadata is static and descriptive of the item as a whole, and does not capture relationships among concepts, nor do article-level subject headings capture all the concepts discussed within the article.

Full-text searching, while allowing searchers to search for ideas mentioned only briefly in an article, suffers from the ambiguities and richness of language, in which a concept, disease, or entity could be described in a number of ways. Neither the metadata of bibliographic citations nor the entire content of full-text articles truly addresses the challenge of increasing both the relevance and recall of search queries while extending the search to disparate types of content.

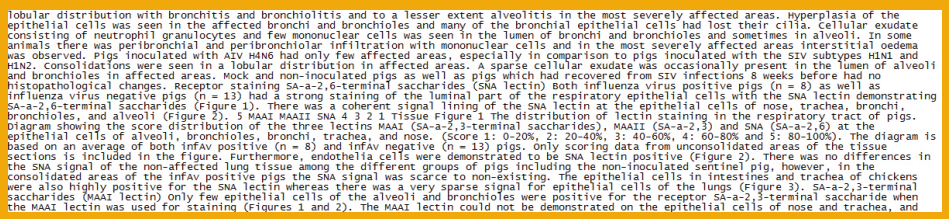
Turbo-charging data through semantic enrichment

One of the areas where information managers can make the biggest contribution is in introducing user groups to the impact of semantic enrichment of internal and external content, and the importance of data cleansing.

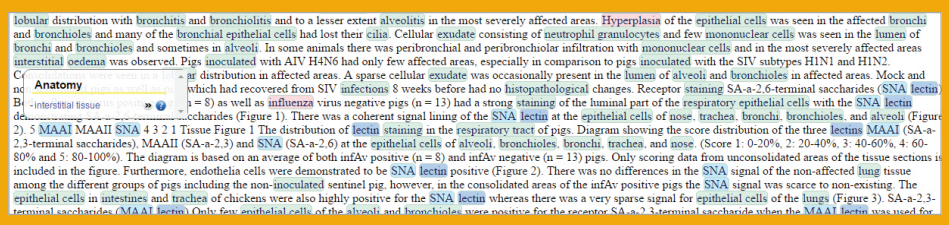
As machine learning and text and data mining initiatives are launched, information managers have an opportunity to serve in consultative roles within their organizations. One of the areas where they can make the biggest contribution is in introducing user groups to the impact of semantic enrichment of internal and external content, and the importance of data cleansing. While researchers are familiar with their field, they may not realize that, without consistent metadata and linkages among domain ontologies, critical data may not be findable.

Machine learning has been hyped as a technology that could soon provide us with an app that takes our symptoms and spits out the appropriate treatment. The reality is that most data a machine learning algorithm would use is not FAIR — it is unstructured, with inconsistently-applied metadata, in a variety of formats, not all of which are fully machine-readable. Concepts may have multiple commonly-used names (amyotrophic lateral sclerosis or motor neurone disease, for example), home-grown or outdated taxonomies may have been used, and varying formats may preclude full analysis. As a result, a machine learning tool that is being trained for pattern recognition with an untreated data set will have less satisfactory results than one that is working with a semantically-enriched data set.

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



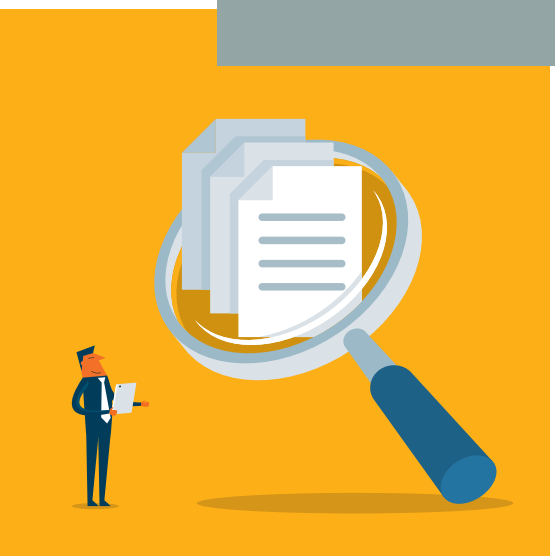
lobular distribution with bronchitis and bronchiolitis and to a lesser extent alveolitis in the most severely affected areas. Hyperplasia of the epithelial cells was seen in the affected bronchi and bronchioles and many of the bronchial epithelial cells had lost their cilia. Cellular exudate consisting 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.
ar distribution in affected areas. A sparse cellular exudate was occasionally present in the lumen of alveoli and bronchioles in affected areas. Mock and
which had recovered from SIV infections 8 weeks before had no histopathological changes. Receptor staining SA-a-2,6-terminal saccharides (SNA lectin)
= 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
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

Using specialized ontologies

Information managers have a number of possible avenues for engagement with user groups involved in using AI technologies for insight. This requires a shift in perspective for information managers, as the focus changes from supporting Boolean searching of a data collection to enabling text mining and increasing discoverability. By using specialized ontologies across both structured and unstructured content, information managers can:

- Add synonyms to queries automatically, ensuring more comprehensive results
- Increase access points to unstructured text and content that is not in an easily-searched format, such as images and charts
- Create links among classes of concepts, such as genes and diseases, so researchers can discover new relationships
- Decrease query abandonment of dissatisfied users by increasing the number of relevant results to accelerate discovery and research

If content is semantically enriched with meaningful metadata and domain-specific ontologies, the information is more actionable across the organization, by a variety of users.



The abandonment of searches is particularly important as younger users expect to use simple keyword or natural language searches in structured databases, just as they query general search engines on a mobile device by speaking their question. Their searches in enterprise information sources only succeed if an information manager has set up domain-specific search filters to enhance precision and recall.

Some text mining projects will mine unstructured content to generate insights from a data collection, but these are one-off benefits, as the underlying content is not modified by the insights derived from the analysis. If, on the other hand, that content is semantically enriched with meaningful metadata and domain-specific ontologies, the information is more actionable across the organization, by a variety of users. While the dataset may have been acquired for a project in one field, adding metadata with links to other enterprise ontologies makes the same dataset valuable to researchers in other fields as well. In other words, enrich once, and benefit many times.

When information managers are brought into these one-off text mining projects to add semantic enrichment to the data, they can offer a unique perspective. They understand how various user groups discover and access content, what external and internal content a project requires, and how to enhance users' queries to better get the insights they need. They understand the wide variety of formats to consider — everything from electronic laboratory notebooks to patent filings, news and press releases, and genomic mapping data — and they know what ontologies to use to add structure to the information.

One challenge information managers often encounter is identifying and accessing ontologies or vocabularies used by a group but not coordinated with other parts of the organization. These may be a locally modified version of a public ontology, a reference database developed by a nonprofit organization or a for-profit company, or a semantic vocabulary developed entirely in-house. Each group may have its own policies regarding who is allowed to contribute changes or who is responsible for keeping the local version reconciled with the original or authoritative version. Unless an information manager or other information professional with an enterprise-wide perspective is managing these local ontologies, the organization loses the ability to leverage the time, expense, and focus required to develop and maintain these resources to benefit the entire enterprise.

While decisions about ontologies are often driven by subject matter experts, information managers can provide centralized management of ontologies and vocabularies within an enterprise, mapping the domain and the different vocabularies being used by each group. Information managers can bring a FAIR mindset to ontology management, by looking at the entire workflow process and identifying points at which data, metadata and tools can be made more findable, accessible, interoperable and reusable, while also being mindful of oversight and authorization concerns.

With a robust ontology management tool, information managers can facilitate collaborative editing of vocabularies, strategic linking of related concepts between ontologies, better interoperability among data sets, consistency of terminology and meaning across projects, and support for multilingual vocabularies. Information managers can identify inconsistencies between ontologies — a particular problem in industries that are experiencing a high level of acquisitions and mergers — and identify new entities or synonyms in rapidly developing fields before third-party ontologies are updated.

> Information professionals must watch — and probe during research requests — for opportunities to explore data analysis needs and concerns, and to look at meeting the user's information needs in a more strategic way that supports deeper analysis and insight.

Four ways to advocate the value of semantic enrichment

Information managers already see the importance and impact of semantic enrichment in search and discovery. However, they may not be involved in the initiatives that require data cleansing and enhancement, as their users may not realize the expertise information managers could bring to the project. Below are four approaches and strategies for raising awareness of both the need for semantic enrichment and the consultative role information managers can play.

When users see the difference in both relevance and recall with a semantically enriched search, they have a better understanding of the kinds of information-related problems that an information center can address.

1

Information professionals must watch — and probe during research requests — for opportunities to explore data analysis needs and concerns, and to look at meeting the user's information needs in a more strategic way that supports deeper analysis and insight. Since users will not ask an information center for a service if they do not think the information center could do the work, information managers must bring up questions of data consistency, discoverability and interoperability with question such as:

- What data are you trying to integrate for this project?
- What formats does it come in?
- Are you able to do the analytics you need to do, or does the data need to be cleansed?

Complex literature search requests can sometimes be indicators a conversation about semantic enrichment would be productive. If the request is "I need a list of all the drugs on the market known as strong inducers of this particular mutation," an information manager could talk about the value of semantic enrichment to support more complex research needs.

2

Information managers can focus on the message that semantic enrichment saves time and delivers better, more relevant search results. Traditional literature reviews require researchers to spend a significant amount of time screening retrieved articles that, while they meet the search criteria, are not relevant. One of the simplest ways to demonstrate the value of semantic enrichment to a data collection is to compare the results of a Boolean search of the text and a search of the semantically enriched collection. While a Boolean search of corticosteroids and asthma, for example, will retrieve thousands of articles that mention both concepts, a query of a semantically enriched database will retrieve only those articles in which the two concepts are related in a specific way.

3

Information managers can use short workshops to promote a semantic enrichment tool; these can be particularly effective if they include examples of use cases and the kinds of outcomes that are only possible through semantic enrichment. As with any effective marketing strategy, the focus needs to be on the benefits of a particular tool to those users' workflow, not the various features available. This may require conversations with individual users to gather examples of how the tool made their job easier or how they could accomplish something new because of the tool.

4

It has always been important for the information center to communicate effectively about the value it provides and impacts it makes to the organization as a whole. An approach information managers can take to address this is to highlight new services that leverage the expertise and resources of the information center. They can reach out to a team or group and offer to develop a customized taxonomy for their data, with domain-specific synonyms and disambiguation of terms. When users see the difference in both relevance and recall with a semantically enriched search, they have a better understanding of the kinds of information-related problems that an information center can address.

This approach is particularly useful in organizations prone to data silos. There are many causes for data being kept by one group and not easily accessible by others—corporate or departmental culture, legacy systems and proprietary formats, and IT budgets that focus on individual departments or functions instead of the enterprise as a whole. While access to some siloed content may need to be limited because of regulatory or licensing restrictions, semantic enrichment may still provide insights from the underlying data to other groups within the organization.

Information management professionals can offer a unique perspective on the value of semantic enrichment and the impact it can have on the enterprise in increased research productivity and improved decision-making. When they provide consultative support to AI initiatives within their organization, they ensure better, more FAIR data and better results.

Mary Ellen Bates

Bates Information Services Inc.

Mary Ellen Bates is the principal of Bates Information Services Inc., providing business insights to strategic decisionmakers and consulting services to the information industry. Mary Ellen worked for over a decade in corporate and government information centers before launching her business in 1991. She received her MLIS from the University of California Berkeley and is based near Boulder, Colorado.

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