Chapter 34

IPM information technology

John K. VanDyk

Today as in no time in history, the management in IPM is supported by information technology. From automated temperature sensors in the field to the desktop computer to the internet website, technology assists us in making pest management decisions. It is therefore necessary to have a framework for thinking about the information flowing through the system and the ways the information is processed. In this chapter we'll focus on a concrete example, namely, that of pest data making its way from the scout or researcher to the decision maker. Along the way we'll encounter common concepts of information flow that are equally useful in other scenarios.

34.1 Information life cycle

Information is all around us, yet we must select which information we want to record. Information has a definite life cycle. It can be recorded, analyzed, summarized, interpreted, shared and finally archived or discarded (Fig. 34.1).

34.1.1 Observations in field or laboratory

The number of insects in a trap, the percent leaf area removed, the damage done to a root system – these are all measurements or observations that represent the foundation of IPM decision making. The grower, researcher or knowledge worker uses the observations or measurements taken to make a statement or recommendation based on analysis. In doing so, the worker draws on past experience, the work of others, and the analysis of many contextual variables such as weather, local landscape, parallel observations at other locations, etc.

34.1.2 Quality control

Peer review or an editor serves as a quality control filter for the information that comes from researchers or models. Only those with the appropriate knowledge and experience are qualified to judge whether the analysis is helpful or not. In practice, when looking at information that makes it into the public eye, this quality control step may be skipped. That is, there is nothing to prevent erroneous information from being published on the internet or sent by email. Consumers of information compensate for this by placing more trust in information sources that have a good history. A good way to think about the review process is that it adds value to information. All other things being equal, reviewed information will be preferred by the information consumer.

34.1.3 Distribution system

After passing quality control, information is presented to a wider audience through print or electronically. We will focus here on electronic

Integrated Pest Management, ed. Edward B. Radcliffe, William D. Hutchison and Rafael E. Cancelado. Published by Cambridge University Press.

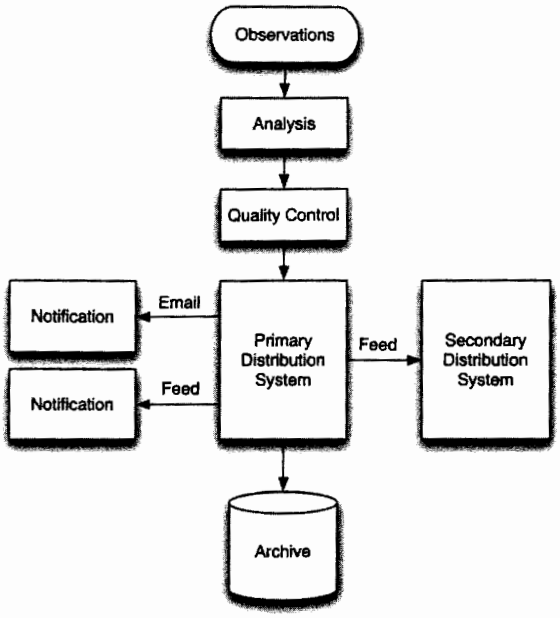

Fig. 34.1 Simplified IPM information life cycle.

information. Information is distributed outward from a trusted source to others. The primary ways to do this electronically are by electronic mail, by creating a page on the World Wide Web, or a combination of these two.

Electronic mail has the advantage of being ubiquitous, but has several disadvantages. First, because electronic mail may be forwarded from person to person, information may be lost along the way. For example, a recipient could truncate a message and remove information about the source of the message in the process. Second, because the authentic original copy of the message is not preserved (except possibly in the sender's outbox), a public reference to the original is not preserved.

Publication on the World Wide Web is fundamentally different than sending information by electronic mail. First, there is one copy, preferably at a permanent uniform resource locator (URL). Second, only the entity that controls the website on which the resource is published has access to change the information.

Primary source

The electronic version of the information being shared should have one permanent online location with a URL. It is helpful if the URL is semantically meaningful; for example the URL *http://example.org/2007/05/23/applemaggot.html* seems to denote some information on apple maggot that was published in May of 2007.

If the information is part of a continuing information series, such as a pest management newsletter, an alias may be created that points the reader to the permanent URL. Suppose a newsletter is published electronically each week. The easily bookmarkable URL *http://newsletter.example.org/latest* might be used for the most recent issue, in which summaries of the five most recent pest recommendations are available. However, a permanent and semantically meaningful URL, such as *http://newsletter.example.org/2007/05* would also be established as each issue is published, and clicking on the title of a recommendation summary would take you to that recommendation's permanent URL, such as *http://newsletter.example.org/2007/05/gypsymoth.html*.

Secondary sources

As information is posted, interested parties need to be notified. Electronic mail is one way to achieve notification, but syndicated web feeds such as RSS (Really Simple Syndication) and Atom (Atom Syndication Format) are increasingly popular. Feeds allow a computer program to check a website for updates periodically. One can think of a feed as broadcasting a channel for other websites or programs to pick up. For example, a second website may present recent headlines from the original website in a sidebar entitled *Recent News*. As new information is posted to the original website, a given headline in the *Recent News* sidebar will move down in the list of headlines and ultimately fall off the bottom of the list. Thus, secondary sources such as the sidebar used in our example point to the primary source but should be considered transient while the original source should be permanent.

34.1.4 Archiving

What happens to old information? Although information that is published at a given URL should remain there permanently so that links to the information do not break, the reality is that websites often disappear or change the URL at which

the information appears. The causes of this phenomenon are numerous: political issues such as funding changes or the renaming or dissolution of the organization that originally published the information; technological issues such as moving to a different platform for serving web pages without preserving the original URLs, and simply bad planning, where pages were not given a permanent URL in the first place.

A good system for presenting information should have organized archives. Failing that, information may be picked up and stored by third-party entities, e.g. *Internet Archive* (2007) and thus remain available after a website has gone offline.

34.2 Adding value

Although information is valuable, it can be made more valuable by adding metadata, sometimes called metainformation. Metadata is data about data. Metadata makes working with information much easier because it allows data to be classified, sorted, indexed and filtered. From an information technology perspective, the usefulness of metadata rests on marriage of the relational model of data management (Codd, 1990) with semantic web concepts such as machine readability and ontology languages (Antoniou & van Harmelen, 2004). For more information about metadata, see *Understanding Metadata* (National Information Standards Organization, 2004).

34.2.1 Taxonomy

Taxonomy is the science of classification. You may be familiar with the classification of biological organisms into groups such as families, genera and species. In the same way, information can be classified, making it easier to discover and retrieve. Information is classified by assigning semantic terms from vocabularies (collections of terms within a semantic group) in a process called tagging.

34.2.2 Tagging

Tagging involves associating a certain word or phrase with a piece of information. For example, an article about the biological control of fire ants might be tagged with such terms as fire ants, *Solenopsis invicta*, decapitating fly, and biological control. Tagging is an expensive proposition, because the person creating the tag must be familiar enough with the problem domain to create correct and appropriate tags. The time of such experts is always in high demand. Typically this cost has prevented the tagging of information. Recently, however, a new approach is being tried. By opening up the tagging of information to the universe of information consumers, the expense of tagging is borne not by the agent creating the information or making it publicly available, but by those who are reading it. This requires a level of trust between information provider and consumer, and risk is involved: the risk that the information may be tagged incorrectly or inappropriately. Still, it is less expensive to review tags that others have created and correct the bad ones than to tag everything correctly in the first place.

Table 34.1 Example of vocabularies and terms

Vocabulary	Terms
Organism	fire ant, decapitating fly
Scientific name	*Solenopsis invicta*, *Pseudacteon tricuspis*
IPM subdiscipline	biological control

34.2.3 Vocabularies

A vocabulary is a selection of related terms. Clearly defined vocabularies are more powerful than simple tagging. The process of tagging described above involves assigning terms to an implicit vocabulary, usually a vocabulary called "Topic" or "Category." Looking more closely at the example of an article on fire ants, we can see that the terms we used to tag the article can be organized into several clearly different vocabularies, as shown in Table 34.1.

Vocabularies can be shared among separate websites if (1) there is agreement on the semantic meaning of the vocabulary and (2) terms in the vocabulary are consistent between websites.

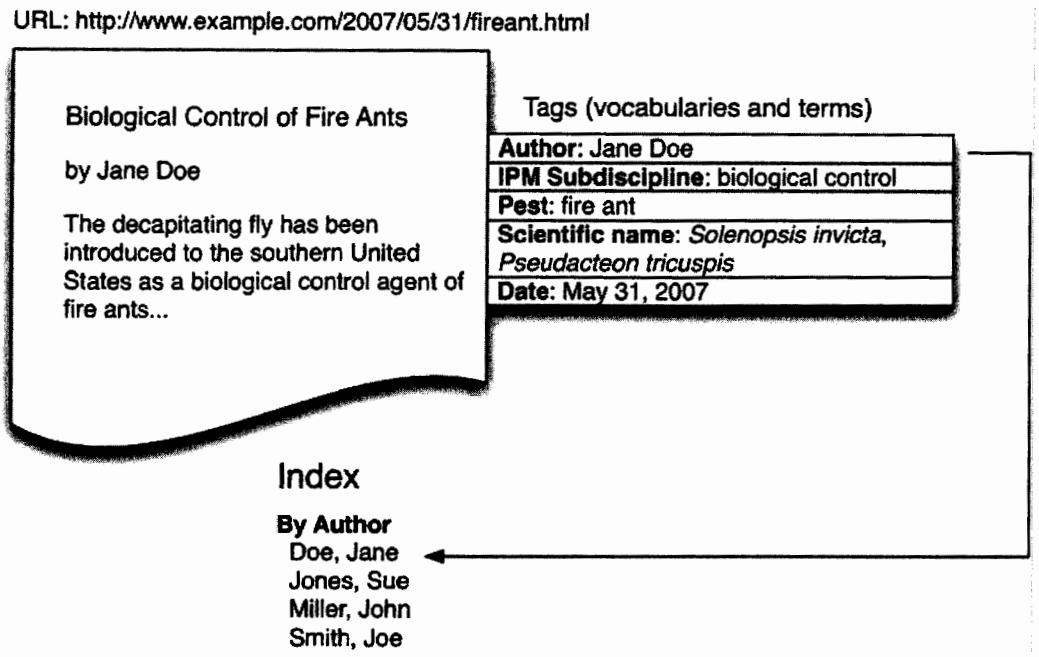

Fig. 34.2 A web page and its associated pieces of metadata, including vocabularies and terms.

34.2.4 Indexing and archiving

The major payoff of establishing terms and vocabularies is indexing; that is being able to browse through information by a vocabulary's terms instead of being limited to performing simple keyword searches. When thinking about IPM information, the vocabularies needed can be identified by asking yourself, how would I like to be able to navigate this information later? Vocabulary names will occur after the word *by*: by pest, by author, by date, by scientific name, by geographic region, etc. (Fig. 34.2).

An even larger benefit would be for providers of IPM information to adopt semantically standardized vocabularies so that information from separate providers could be merged together and browsed as one body of information. However, at the present time there is no standardized set of vocabularies for IPM information. There are several reasons for this. First, it's difficult to agree on standards because information can be classified in multiple ways. This is true even in small, clear-cut sets of information. Second, integrated pest information is by definition quite diverse. It encompasses multiple disciplines from agricultural engineering to zoology. Establishing vocabularies across disciplines is very difficult. That leaves vocabularies that are already very standardized, such as geographic information and scientific nomenclature.

34.2.5 Term relationships

Another benefit of providing metadata with information is the ability to find contextual information quickly. The algorithm for doing so is remarkably simple. If all information is tagged, we simply search through the information for other information that matches the term with which the current information is tagged. For example, we might have an article about Asian soybean rust (*Phakopsora pachyrhizi*) in an online publication dealing with plant diseases. We can then search for other articles with that same tag. If named vocabularies have been used instead of the implicit *Topic* vocabulary used by general tagging, the specificity of the related articles can be improved; that is, searching for information inside a *Plant Diseases* vocabulary for the term Asian soybean rust will give more specific information than a general search for that term.

34.2.6 Filtering

Filtering is another benefit of tagging. Filtering allows information to be discarded based on a

vocabulary–term pair. For example, one might want to see all of the information on the effects of *Pythium* infection on maize. Because *Pythium* affects both maize and soybeans, a simple search for information tagged with the term *Pythium* will not be especially helpful. What would be helpful would be two vocabularies, a *Crop* vocabulary and a *Scientific Name* vocabulary. Then a compound search could be performed resulting in only information where the *Crop* vocabulary contains the term maize and the *Scientific Name* vocabulary contains the term *Pythium*.

34.3 | Aggregation

Aggregation is a simple but powerful concept. Aggregation means taking data from disparate sources and merging them into one data stream. In order to join the data together, all sources of data must adhere to a common schema; i.e. a number of data fields that have the same semantics. For example, if scouting data are being aggregated from a number of sources, the same units of measurement (e.g. insects per plant versus insects per leaf) and geographic information should be used by all participants.

34.3.1 Pest monitoring

A common theme in IPM is the monitoring of insect populations across a geographic area. In the upper midwestern USA, black cutworm (*Agrotis ipsilon*) is a pest of maize that does not overwinter there but invades each year from overwintering sites in the South. The invasion is tracked by observing pheromone traps and aggregating the trap counts to get a picture of what is happening in the field; the aggregated data can then be visualized and used for predicting a timeline for damage (Rice & Pope, 2007). Thus, aggregation is a way to get an overview or summary of many separate events.

34.3.2 Feeds

In the same way that aggregation of pest monitoring data gives a summary of what is happening in the field, aggregation of IPM information itself can give a summary of what is happening across informational boundaries. This is accomplished through the use of open standards for sharing information and the software programs called *news aggregators*. Such programs are also referred to as *feed readers* or simply *aggregators*. They work by polling special files known as *feeds* provided by websites. The presence of a feed on a website is usually indicated by an icon as well as the presence of some additional code in the website's hypertext markup language. A website can provide multiple feeds. For example, the website may provide a main feed that contains updates of any page on the site; as new pages are posted they appear on this feed. For a website that provided IPM news, a new article about the presence of black cutworms in pheromone traps would appear on the feed, and those people using aggregators to track news across multiple websites would receive an indication that a new page had been posted. The notification is usually in the form of the title and a brief synopsis of the article.

If the information on the website has been properly categorized into vocabularies and terms, the website may expose feeds for individual terms. The new article on black cutworms would then appear not only in the main feed but also in the feed for the Black Cutworm term of the *Pest* vocabulary, or the *Agrotis ipsilon* (Hufnagel) term of the *Scientific Name* vocabulary.

34.4 | Feedback

The information system described above offers information in itself. One can make observations about what information is being viewed most often, what information is seasonally accessed and what terms in a vocabulary garner the most attention. These observations are obtainable from access records in web server logs.

Also, the presence of terms in vocabularies can be used to assess the breadth and depth of coverage a website has. If all information has been tagged, it becomes easier to identify areas of strength and weakness in coverage. The logs generated from the use of a search feature on a website are very helpful, as they generally indicate what is being searched for. A low level of correlation between keywords being searched for in a site's search engine and the terms in the site's

vocabularies may indicate that the information presented on the site is not what the site's users are expecting to find. Such observations and following analysis can be used to reorient the site to be more useful.

34.5 Conclusions

The use of information technology to obtain and manage IPM information will continue to grow. By applying the basic principles of information taxonomies such as tagging information with terms from vocabularies, filtering and aggregation, knowledge workers will have the necessary tools to become increasingly informed about the realm of IPM.

References

Antoniou, G. & van Harmelen, F. (2004). *A Semantic Web Primer*. Cambridge, MA: MIT Press.

Codd, E. F. (1990). *The Relational Model for Database Management*, v. 2. Reading, MA: Addison Wesley.

Internet Archive (2007). *Internet Archive*. San Francisco, CA: Internet Archive. Available at www.archive.org/.

National Information Standards Organization (2004). *Understanding Metadata*. Bethesda, MD: National Information Standards Organization. Available at www.niso.org/standards/resources/UnderstandingMetadata.pdf.

Rice, M. E. & Pope, R. E. (2007) *Integrated Crop Management Newsletter*. Ames, IA: Iowa State University. Available at www.ipm.iastate.edu/ipm/icm/2007/5-7/blackcutworm.html.