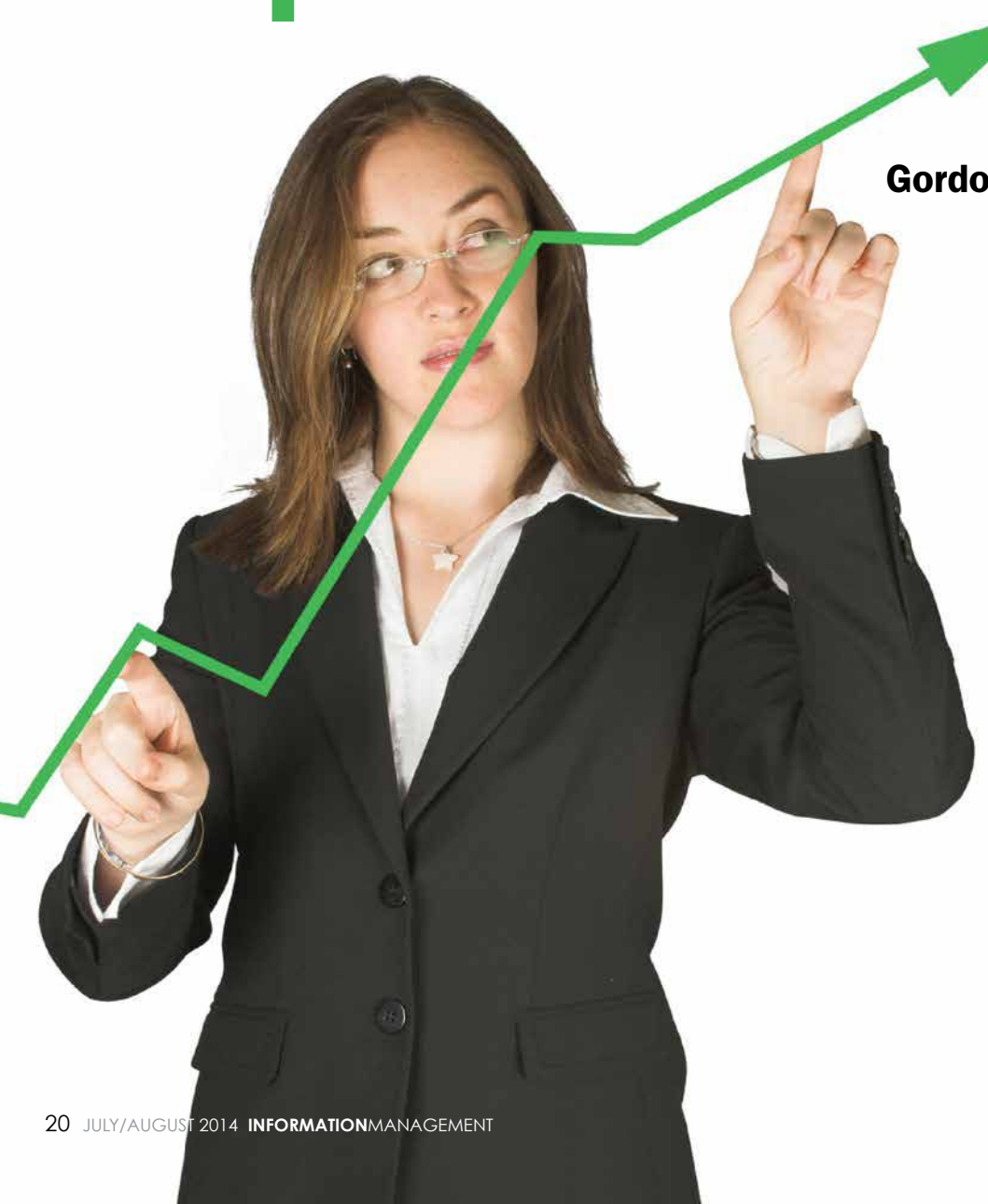
Top IG Tech Trends: Auto-

Gordon E.J. Hoke, IGP, CRM



Classification & Big Data

Information governance (IG) is an emerging practice in several disciplines. Law, records and information management (RIM), information technology (IT), and others define it from their own perspectives. For example, some attorneys equate IG with defensible disposal, while some technologists see it as a storage or architecture issue.

April's Executive Conference on Information Governance, co-presented by ARMA International and The Sedona Conference® (TSC) and attended by more than 100 people from at least a half-dozen disciplines, did not seek or achieve a consensus definition, but the shared perspectives did encourage a cross-fertilization of ideas.

ARMA International defines IG as:

> A strategic framework composed of standards, processes, roles, and metrics that hold organizations and individuals accountable to create, organize, secure, maintain, use, and dispose of information in ways that align and contribute to the organization's goals.

This definition suggests IG is actionable – a strategy for accomplishing goals. In contrast, TSC's definition takes a descriptive approach:

> An organization's coordinated, inter-disciplinary approach to satisfying information compliance requirements and managing information risks while optimizing information value. As such, Information Governance encompasses and reconciles the various legal and compliance requirements and risks addressed by different information-focused disciplines, such as RIM, privacy, information security, and e-discovery.

The differences need to be acknowledged. Full benefit from IG requires an appreciation of both the descriptive qualities and the functional contributions. For example, TSC emphasizes risks twice while risk is only implicit in ARMA's definition. In contrast to TSC, ARMA emphasizes comprehensive precision. Until consensus emerges, practitioners will benefit from applying both definitions and keeping in mind the perspectives of others.

IG Stakeholders

The varied definitions notwithstanding, consensus was alive and well at the executive conference. For example, no voice contested the supposition that success in any IG initiative requires a strong executive champion. Further consensus ascribed IG success to overcoming the separation or isolation of such departmental stakeholders as IT, finance, RIM, legal, research and development, accounting, sales, human resources, procurement, and others. In many organizations, these departments function as what conference presenters termed "silos."

IG offers value: each functional area stands to benefit from harvesting synergies. Better coordination leads to less redundancy with better operations and compliance. Leaders from these areas (siloed or integrated) are stakeholders in IG. They stand to benefit from ending departmental insulation. Indeed, the task of the IG professional is to facilitate and enable cooperation. According to one session leader, the silo effect may be emotional as well as operational; another conference presenter invoked "organizational psychology" as a useful tool for IG.

IG Technology Challenges

For successful IG, participants must be able to share information. Their information systems need interoperability or, minimally, communication links. Further, technology should enhance operational efficiencies and

The first **Executive Conference on Information Governance** convened in April on Amelia Island, Florida. Presented jointly by ARMA International and The Sedona Conference® (TSC), the event featured plenary sessions with multiple presenters.

Guidelines for presenters and responders required "dialogue, not debate," a motif common to TSC meetings. Leaders encouraged attendees to comment on the conference in social media and more traditional forums, and many posted on Twitter. In the interest of free expression and exchange, however, conference guidelines required that no direct quotes appear and ideas not be attributed to an identifiable individual.

This report respects those requests and focuses on technological facets of information governance that surfaced at the conference. Definitions and parameters are presented for context.

The adoption of auto-classification appears to be more a matter of timing, cost, and tools rather than whether it will become a norm.

facilitate synergies. Clearly, an executive champion, council of stakeholders, and departmental implementers are essential, but unless technology plays a robust role, only limited IG progress will accrue from policies, procedures, and practices.

An information architecture can either help or frustrate IG efforts, but no single information model is ideal. Centralized or distributed servers may suffice, although the derivative issues vary. IG can thrive inside a tight firewall or by employing a public cloud; these are questions of style, not adequacy. In any situation, an enabling architecture facilitates IG stakeholders' collaboration.

There are hardware considerations as well. Some vendors' IG tools require a huge number of information processing cycles. Many require increased network bandwidth. Distributed architectures require high-speed communication. Organizations considering IG-enabling software may find themselves looking at estimable hardware purchases. Hosted solutions may mitigate this need.

Similarly, systems and applications play a key role. Legacy databases may resist IG. Migrating old systems to archives or newer, full-featured systems is costly and may risk the integrity of data and metadata.

E-mail used as a records repository is problematic for many organizations. SharePoint is another common conundrum: relatively few organizations govern it comprehensively, and records management is not the platform's forte.

IG Technology Trends

A half-dozen vendors and consultants at the executive conference offered professional solutions that are on the market and need evaluation for individual organizational needs.

Perhaps the good news for the technical side of IG is that the field is immature, meaning that the vendors have varied approaches. This lack of standardization suggests that any inquiring IT group may well find an approach that is compatible with its unique needs.

Within this context, two technology trends dominated presentations and conversation at the event. One, auto-classification, comes under many names – some related to its use. The other major trend, appearing under the umbrella of big data, depends on the efficacy of auto-classification.

Auto-classification

Synonyms or near-synonyms of auto-classification include automated e-discovery, predictive coding, content analysis (or analytics), and computer-assisted review. The software products sold with these names are tuned to somewhat different functions. Their approaches differ and certainly their underlying algorithms are distinct. Their commonality: they use machines to make valuable information more accessible and useful, and under most conditions they do it more quickly and accurately than humans.

In RIM, auto-classification arose in the last decade. Practitioners knew that some human record owners were neither quick nor accurate in declaring information as records and assigning them to records series with disposal dates. Some (at the time) brazen software developers suggested that their algorithms could declare records better than human workers could.

As early as 2007, software developers pointed out that a human that could assign 90% of appropriate records to the right record series 90% of the time had roughly the same success rate as a computer that could consider and assign the right series to *all* records 80% of the time. Because the machine was faster, though, they gave the advantage to the computer.

Over the intervening years, algorithms improved, processing speeds rose, and developers touted computer classification accuracy in the 90% range. Commensurately, the amount of captured information and records rose precipitously, to the point where humans could not expect to keep up without automation. This same phenomenon led to big data, which is discussed below.

In the legal arena, emerging case law and amendments to the Federal Rules of Civil Procedure gave parameters for court-admissible electronic information. This admissibility expanded the scope and the significance of electronically stored information. Manual inspection of large numbers of electronic records by high-priced law firm staffs raised the stakes. The software that could quickly analyze digital records – which previously was too expensive – became cost-effective. A new industry arose around legal search, e-discovery, and computer-assisted review.

A related technology that developed simultaneously was content analytics. This technology moved beyond using metadata and keyword search tools to identify and classify records; content analytics can actually recognize the *meaning* contained in text.

Language is highly complex, and for machines to rec-

ognize and flag relevant text based on its meaning, there must be sophistication in software and power in hardware. Consider two e-mail messages. The first says:

> *Dear Bill,*
> *You and Sue are invited to a barbecue on our new patio Saturday. The contractor did a great job.*

The second says:

> *Dear contractor:*
> *The work you did was terrible and your bill is invalid. I may sue you for damage you did.*

Some of the words are the same, but the meanings are very different. For a machine to recognize the difference, it needs profound algorithms that go beyond the dictionary definitions of the words, extracting meaning from the context and syntax – that is, the way the words are used.

Auto-classification and its variations have improved significantly in the last several years, and forthcoming versions should be even better.

Regarding auto-classification, executive conference participants fell into these general categories:

- *True believers*, who see the new tools as the only realistic way to bring the risks of undeclared records, misapplied records series codes, and unfound records down to an acceptable level at an acceptable cost
- *Skeptics*, who fear the consequences of relying on immature technology
- *Practitioners,* who appreciate the capabilities of auto-classification but do not see a practical way to implement it due to such things as limited budgets, technical expertise, user acceptance, and staff resources

In any case, a clear trend is the continued evolution of auto-classification. In the face of ever-increasing quantities of electronic records, the adoption of auto-classification appears to be more a matter of timing, cost, and tools rather than whether it will become a norm.

Big Data

The aforementioned rapid growth of electronic information and increase in the volume of records and court-acceptable information lead to big data.

Executive conference presenters vehemently contested the common misconception that big data is just "more of the same." The amount of data available for recordkeeping has grown exponentially over the last few years, and indications are that the rate of growth will continue. Not all organizations create or use big data, but those that do soon realize that techniques for processing the onslaught of information are discontinuous with earlier ways. The management tools are different as well.

Although big data was described as early as 2001, the executive conference offered a 2013 definition from the nonprofit association ISACA: "Data sets that are too large or too fast-changing to be analyzed using traditional relational or multidimensional database techniques or conventional software tools to capture, manage, and process the data at a reasonable elapsed time."

Where does big data originate? Why is there so much of it? Despite the many potential answers, two illustrations suggest some sources:

1. *The World Wide Web and mobile applications.* Any number of website owners are intensely interested in the behavior of their site visitors. They record every mouse or keyboard click and

every screen touch. The number of data points occurring at a popular website can be enormous. Similarly, mobile applications' owners record the activity of their users. While the reasons owners collect this information vary, the volume of data amounts to exabytes.

2. *"The Internet of Things."* Tens of billions of devices have sensors or signal relays that connect to the Internet. These range from radio frequency identification (RFID) chips reporting locations or inventories to sensors along railroad tracks that recognize and report boxcar wheels with hotboxes. In a home example, refrigerators may have sensors that report to the owner's mobile device if the internal temperature rises above a set level.

Big data is significant for more than its tremendous volume. Algorithms can organize this information into meaningful, predictive patterns. For example, Amazon knows that a specific percentage of its website visitors that looks at a book will later buy it.

Amazon also knows the geographical location of the viewers. With this information, the retailer ships an appropriate number of copies of a particular title to warehouses near its viewers, even before a viewer turns into a buyer. This facilitates the quick delivery that engenders customer satisfaction.

In another example cited at the executive conference, a father learned his teenage daughter was pregnant because a retailer – predicting behavior based on web views and/or store movement – began sending direct mail advertising for baby products.

Similarly, analysis of big data can affect momentous events such as natural disasters and terrorist activity. Big data analysis enables severe weather alerts, and the U.S. National Security Administration uses big data to predict enemy strikes.

Applying ethics to big data use is only beginning. While identifying and anticipating equipment failure are straightforward, collecting and acting upon information about people present moral and legal dilemmas. The questions may fall into three categories:

1. Personal information that individuals knowingly and freely provide to data collectors (for example, during registration at a website) for a known and approved use
2. Personal information collected about individuals without their knowledge or specific permission, such as location and movements obtained through cell phones
3. Personal information individuals knowingly and freely provide to data collectors that is analyzed for additional meaning (sometimes paired with external data) and used for purposes beyond the intent of the original permissions

Each of these uses of big data carries ethical implications. The ethical codes of attorneys and records managers certainly do not extend to all big data users, and the right path may not always be clear.

The Twain Shall Meet

The Generally Accepted Recordkeeping Principles® (Principles) apply to big data. The Principles rise above volume, source, medium, speed, and other variations. Just as they provide a comprehensive governing framework for both paper records and digital images, they guide the management of information in databases and big data repositories.

While the Principles apply universally, the methods and techniques for applying them vary by the nature of the information's attributes. This is where auto-classification, predictive coding, and content analytics meet big data. Since the quantity of big data is, by definition, too large for conventional database entry and processing, powerful computers running advanced algorithms are the tools of choice for big data governance. These algorithms and related policies can apply the Principles, especially retention, availability, protection, and disposition.

In the evolution of technology, capabilities typically come first, while governance, controls, and ethics arrive later. This is the case with big data. Years ago, users began exploiting big data, but attorneys at the executive conference reported that litigation based on big data has come to courts only recently. To prosecute, defend, and argue these cases, traditional discovery methods are impractical. Effective research requires computer-assisted review and other automated tools.

Records managers will similarly find these tools indispensable. They provide more than automatic records declaration. They can apply and release legal holds. They can protect records from unauthorized access. And of similarly vital importance, they can auto-delete records when retention periods are completed.

Looking Ahead

The executive conference received many positive evaluations, and plans are in motion for a 2015 edition. Undoubtedly IG technology and its rate of use will continue to evolve in the coming year. The functional and ethical challenges will grow as well.

Facing burgeoning volumes of information, practitioners will be hard-pressed to maintain current rates of success. Progress may well depend on leaders' ability to harvest synergies from inter-disciplinary collaboration. **END**

Gordon E.J. Hoke, IGP, CRM can be contacted at ghoke@mindspring.com. *His bio is on page 47.*