Six Steps for Creating a 'Super Data Map'

Creating a "super data map" that not only captures metadata about where and in what media information resides, how it is used, and who owns and has access to it, but also integrates legal, compliance, privacy, and IT attributes along with a record retention schedule, can lower risks, reduce costs, and be easier to maintain than separate, single-purpose databases.

Mark Diamond

Do you know where your records actually live – in which systems and on what media? How about your privacy information? Do you know what content is where when you need to place a legal hold?

Multiple groups in an organization need to know what information lives where for a number of purposes. These groups, including legal, IT, and records and information management (RIM) professionals, often take disparate approaches to identifying and classifying the same information, multiplying the work and producing a variety of results.

Organizations that want to link retention schedules and policies to repositories have an even more difficult task. Extending a records retention schedule to capture other types of metadata, such as privacy and security fields or pointers to systems of records, quickly can become overwhelming and unmanageable. What's needed is a better approach. It's time to create a data map.

Defining 'Data Map'

A *data map* is a database that captures an inventory of what you have, where it is, and who is responsible for managing it. It can track record types, personal and confidential data classifications, documents and other types of paper and electronically stored information (ESI), and key metadata, such as how it's used, for what purposes, and who has access to it.

Data maps can track information across a variety of media, systems, and locations. Because information and data are continually created, deleted, and moved, an effective data map is dynamic and updated regularly. Maintaining it is a great challenge, but good map design can make it much easier.

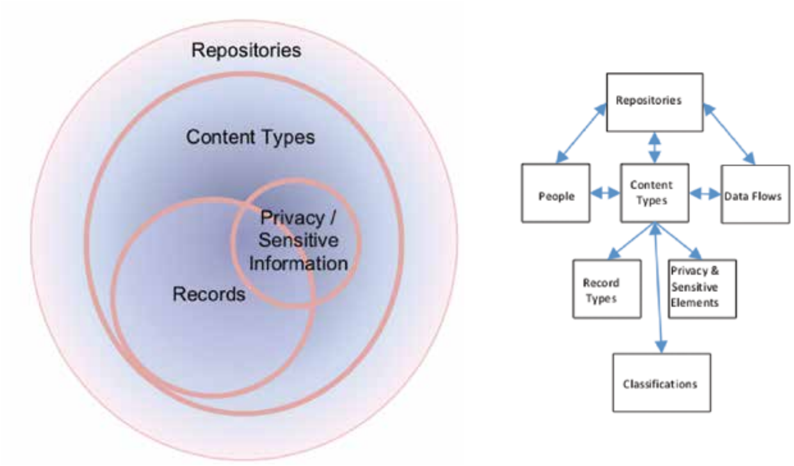

Figure 1: Super Data Map: Fitting the Pieces Together

Identifying Users

A number of business functions need to track the location of documents and data. These include the following:

Application and Infrastructure Management

IT groups need to catalog enterprise applications, repositories, and systems across the organization. Such information helps guide backup and archival strategies, disaster recovery plans, and capital spending.

RIM

RIM professionals need to know which records reside in which repositories, track systems of record, identify what records are convenience copies, and manage retention requirements. They also need to identify and defensibly dispose of expired, duplicative, and low-value data and documents.

Legal and Compliance

Litigators and investigators need to know the location of ESI and hardcopy content that may be relevant in a legal proceeding or investigation. This knowledge enables them to issue narrower legal holds, thereby reducing the costs of discovery and increasing defensibility.

Legal and compliance teams need to track trade secrets, intellectual property, and other kinds of private and confidential data. They also have to ensure that employees, customers, and other legitimate stakeholders have access to data, while unauthorized or non-legitimate users don't.

Auditors need to track financial and compliance information that is relevant to one or more specific regulations, including the Sarbanes-Oxley Act of 2002, the Foreign Corrupt Practices Act of 1977, and others.

Privacy

Privacy professionals have regulatory and statutory requirements to identify and track personally identifiable information (PII), protected health information (PHI), and other privacy data. This may also include privacy data flows.

While the needs for mapping vary across functions, the mapping process is very similar. Creating a single, "super" data map that combines records, privacy, discovery, and other drivers and serves multiple masters is easier, more efficient, and costs less than building and maintaining multiple maps.

Defining 'Super' Data Map

As shown in Figure 1, a *super data map* identifies the repositories, applications, and storage locations where information can live. Within the repositories are *content types*, which are discrete documents, databases, images, and other content that must be managed for retention or security. Important subsets of the various content types are business *records*, which carry a mandated retention period. Private and sensitive information may be regarded as content types or records, depending on the level of detail to which they must be managed.

Creating a Super Data Map

At minimum, the map includes descriptions of applications and systems; types of *unstructured content* (e.g., documents and images) and *structured data* (e.g., database elements) included in each; the sources and locations of data; and the involved personnel (business and IT custodians). If created in a relational database, super data maps also can incorporate record retention schedules and data security classification policies, providing one place to track data and repositories and linking this information to relevant policies.

Sample Fields for System Information in Data Map

System Name	The system, application, or repository where data is stored
Description	A brief description of the specified system/repository, which may include information about the primary users and the type of data stored there
Hosted	Indicates whether the application is hosted internally or outside the organization
Status	The status of the specified system/repository as of the "Last Update" date. A drop-down menu provides "Current" or "Retired" options.
Roll-Out Date	The first date on which the specified system/repository was available to store data
Retirement Date	For inactive or legacy systems/applications/ data storage locations, the last date on which the specified system/repository was actively accepting new data
Data Structure	Identifies the type of information housed in the specified system/repository. Standard descriptions include unstructured (e.g., flat files saved on the network), semi-structured (e.g., MS Outlook e-mail), and structured (e.g., database records from applications such as PeopleSoft or Oracle).
System of Record	Identifies whether the repository is considered a system of record or a secondary or reference source
Information Classes	Lists any information, content, or record classes that may be contained within the system/repository (used as a single point of collection to aid in more granular linking of records/information classes to systems/repositories)
PPI Sensitive Info	Indicates through a "Yes," "No," "Maybe" drop-down menu whether personal protected information (PPI) or other sensitive information exists in the repository (used to flag repositories with PPI or sensitive data to allow for more granular linking or classification as appropriate)
Custodians	Lists the name(s) of key business, legal, and IT contacts or business unit subject matter experts with ownership, responsibility, or knowledge of the system/repository
Retention	**Retention Backup** – Describes the current back-up system for the repository, including frequency, media type, location(s) of backup media, etc. **Retention Policy** – Indicates how long information should be retained in the repository

Table 1: Sample Fields for System Information in Data Map. You will want to customize the fields to your needs.

Following are six steps for creating and maintaining a super data map.

1. Form a Cross-Functional Committee

An important success factor for a data mapping project is the formation of a cross-functional team to oversee the effort. The team should include key stakeholders from legal, RIM, and IT, as well as end-users from business units, who have the best understanding of how information flows through and outside the organization. Once the stakeholder groups understand the challenges at hand and the "win" in it for them, they'll be willing to participate, ensuring a map that is usable across the organization.

2. Gather Input from Stakeholders

A super data map will succeed – and scale to meet future needs – if the business requirements are well-defined and agreed-to across the organization early-on. Ask committee members:

Which constituencies will use the data map, and how will they populate and consume the information? Including two or three functions can meet the needs of many.

Will the map serve just one or many purposes? The trick is to make the map useful for any given function without getting too detailed and overwhelming the structure. When it doubt, keep it simple.

What data elements will be collected and maintained for these repositories (e.g., application names, record types, custodians, server locations, backup methods, storage size, format)? Use the answers to create an in-depth database table for each repository that contains detailed content types, as well as additional reporting capabilities to allow production by content type. This allows users to search on specific content types to find the associated repositories. Don't get too detailed, though. For example, a data map may identify that purchasing records live in a specific place, but it should not be so detailed that it shows where contract negotiations from Customer ABC live.

Will the map track privacy information at the object level (typical files or database records) or at the element level (fields within an object)?

How many repositories will the map address? While an enterprise may have hundreds of repositories, 80% of the relevant information may live in just 20; start with these first.

Are there limitations as to the accessibility of information? Inaccessible repositories might include those created or used by electronic media no longer in use, redundant electronic storage media such as backup tapes, or those from which retrieval involves substantial cost.

Each stakeholder group has a unique perspective and a list of what it wants to be included in the map. But, including too many fields and discrete data points will lengthen the collection process and make it difficult to maintain the map.

Conversely, if the scope is too narrow, important data points could be missed, resulting in an ineffective map and the need to re-collect data. The key to good data map design is balance and tolerance of the imperfect; it will be a trade-off among comprehensive data collection, maintainability, and ease of use.

Start with a pilot or trial version of the data map, populating only a sub-section before collecting data on a large scale. Build and improve the map through iteration, as the requirements of multiple groups and the significance of additional repositories and content types are identified. This process will test the structure, allowing early assessment and adjustments to be made and resulting in the proper balance for the data map design.

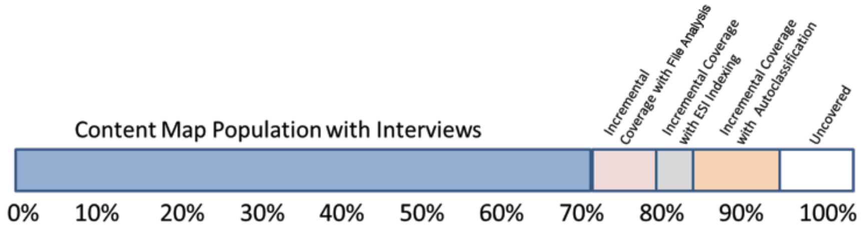

Figure 2: Average Percentage of Data Collected by Collection Method

Table 1 provides an example of the types of attributes that might be tracked within a data map. The actual fields to be included, though, will be dependent on the organization.

3. Choose the Right Structure

Picking the right tool to house your data map is important. There are three options:

MS Word or Excel. These programs may be suitable for retention schedules or very small data maps, but quickly become overwhelmed due to the many-to-many interelationships between the data elements.

MS Access or SQL Server. A simple-to-use but fully functional relational database can be ideal. When designed well, they are capable of mapping significant amounts of data.

Commerial Software. Some very large organizations may wish to keep their data maps maintained through direct links from other applications, such as the HR module from an enterprise resource planning (ERP) system. In these specialized cases, organizations may want to consider purchasing a commercial software tool to hold the data map. The drawback, however, is that these tools may be difficult to customize for specific use cases and environments.

4. Collect Data to Populate the Map

Populating the data map means creating for each repository an in-depth database table entry that contains content type details and creates capabilities for reporting by content type. This framework allows stakeholders to search on specific types of content relevant to their respective use-case and find the associated repositories and other important data elements.

But before the data map can be populated, information must be collected. Following are three of the best approaches for collecting information.

Interviews. Interviewing a cross-section of employees is surprisingly effective. They provide useful guidance when the data to be collected is well-structured (i.e., are of a specified format and can be easily described) and when stakeholder behavior can be categorized (i.e., the expectations of individual groups can be clearly articulated).

Surveys typically miss nuance, such as the pain people may feel when dealing with particular systems and kinds of information. Individual and small-group interviews can uncover real issues and challenges that simple, form-oriented surveys often miss. In practice, surveys followed up with interviews provide excellent guidance and insight.

ESI Scanning and Keyword Index Tools. Automated tools can sort through and index huge volumes of information, making it easier to inventory and classify data. Rules-based approaches use keywords and synonyms along with Boolean logic that

…no automated technology can, by itself, point at a collection of information and then define and populate a data map in a way that is defensible and comprehensive.

is often associated with search engines to confirm objectively a category match with a content item. The precision and completeness of rules-based systems are good when the information to be classified contains sufficient metadata and/or keywords.

Predictive coding goes farther than rules-based systems. This machine learning approach uses established statistical models and a set of keyword-rich "exemplar" documents to train the software about the context and meaning of information. With predictive coding, relevant information can be identified for each concept in the category scheme. This is especially useful when there is not enough metadata available or when large collections of information are spread across multiple data sources, such as e-mail, SharePoint, and file shares (i.e., content "in the wild").

Autoclassification. Originally intended to improve the consistency and accuracy of records categorization, autoclassification software can be suitable for locating many types of documents and files – especially when such items are already housed in supported document management systems and repositories – and can make information easier to search and retrieve. As with predictive coding, autoclassification software requires considerable up-front manual effort and system training.

Automated tools have become sufficiently trustworthy to assist humans in their decisions or, in some cases, to supplant human intervention. The suitability of a particular technology depends on the volume of information to be reviewed, the desired accuracy of the results, and the amount of manual effort and expense that an organization is willing to invest. See figure 2.

At this time, no automated technology can, by itself, point at a collection of information and then define and populate a data map in a way that is defensible and comprehensive. And, none of these tools can establish how the information got to where it is or how to remediate problems. Manual effort is also required.

5. Integrate Retention, Security

The same relational database used to house your data map can also hold your records retention schedule. Furthermore, since repositories are managed as separate elements in the map, creating linkages between record types and their respective repositories is straightforward.

This also applies to data security classification for privacy and other sensitive information. Mapping security levels to elements within a repository allows for easier execution of security policies and provides a convenient view of what sensitive information lives in each repository.

The complexity of the data map increases through embedding schedules and policies, so keep in mind the importance of keeping it simple. Well-thought-out and well-designed map taxonomies – with a preference for simpler – yield benefits.

6. Maintain the Map

As new applications, repositories, and tools are introduced, the information contained in the map can become obsolete; on average, a well-designed map will experience about 20% data "drift" per year. Accountability for ongoing maintenance should be spelled out from the beginning of the project. Identify the responsible parties and the appropriate procedures to be used (e.g., interviews and surveys), and train staff on processes and maintenance. Those responsible for maintaining the map must do the following.

Incorporate IT system change management procedures. Every time IT commissions or decommissions a system or repository, part of the IT system change management process should be to update the data map. Doing so will often address the majority of the changes in the environment.

Leverage discovery to feed the map. New and ongoing litigation will uncover unexpected sources of information that are subject to discovery. Feed information gleaned from the discovery process to update the map.

Develop a regular refresh process. Beyond depending on IT system change management and e-discovery, organizations may want to refresh their maps every 12 to 18 months through the same processes used to initially populate the map. Map maintenance is typically less difficult and much faster than the initial map generation since it will be focusing only on changes.

As is true for developing the map, maintaining the map is best done as a shared process by multiple stakeholders. Many functional hands make for lighter map maintenance work.

Sharing Final Words of Advice

A good super data map can be a boon for RIM, e-discovery, privacy, compliance, and IT. It is an essential navigational tool for climbing the information governance mountain.

So, invest the time needed to design a map that matches your organization's needs. It will pay off with its ease of use and maintenance. Take a balanced approach and include multiple stakeholders. Walk before you run; build the map through iteration, tackling the most relevant repositories first, then working down the list. And don't let perfect be the enemy of good. **END**

Mark Diamond can be contacted at mdiamond@contoural.com. *See his bio on page 47.*