Information Interoperability and Provenance for Emergency Preparedness and Response

Len Seligman, Barbara Blaustein, Peter Mork, Ken Smith, Neal Rothleder
{seligman, bblaustein, pmork, kps, neal}@mitre.org
The MITRE Corporation, McLean, VA

Abstract—*Improved situation awareness is a key enabler of better emergency preparedness and response (EP&R). This paper describes two important challenges: information interoperability and provenance. The former enables meaningful information exchange across separately developed systems, while the latter gives users context that helps them interpret shared information and make trust decisions. We present applied research in information interoperability and provenance, collaborations with leading industrial and academic partners, and illustrate how our tools improve information sharing during preparation, training/exercises, ongoing operations, and response.*

1. INTRODUCTION

To better support a wide range of homeland security missions, diverse organizations need to more quickly share, fuse, and make sense of information from far-flung sources. A key barrier is the cost and time required to achieve information interoperability—i.e., meaningful information exchange among separately developed systems. Industry and academia are building tools that address pieces of the problem, yet these tools are often expensive, and they do not interoperate with other vendors' information interoperability tools. Government agencies need powerful, affordable, vendor-neutral interoperability tools to achieve more agile information sharing. In addition, when information is combined from unfamiliar sources, users have difficulty interpreting the data and knowing whether to trust it. Currently, there are few practical tools to help users understand information provenance (also known as lineage or pedigree); that is, where did this information come from and what processes were used to produce it?

This paper describes the importance of information interoperability and provenance to one critical homeland security mission, emergency preparedness and response (EP&R), and presents applied research at The MITRE Corporation that addresses critical gaps. We then illustrate how the resulting emerging capabilities contribute to enhanced situational awareness in an emergency and how they would play out during preparation, training/exercises, ongoing operations, and in response to an actual emergency.

We begin by describing the challenges of information interoperability and provenance in more detail and then describe how addressing those challenges benefits EP&R.

1.1. Information Interoperability

The goal of information interoperability is to make available information that sources have and are willing to export and to make it understandable to consumers. We decompose the challenges into three levels.

Level 1: Overcome geographic distribution and infrastructure heterogeneity. Data can be widely distributed geographically. In addition, to access the data you must overcome several types of infrastructure heterogeneity including:

- Different data-structuring primitives, such as relational database tables versus XML versus objects
- Different data manipulation languages (such as SQL or XQuery), proprietary data languages, and sources with no query language that require use of a general purpose programming language (such as Java)
- Different platforms, operating systems, networks, etc.

Level 1 challenges are not as resource-consuming as the others because Internet technologies and off-the-shelf products handle most of the challenges. In certain

environments (such as message exchange across military and civilian agencies), however, significant engineering is still required at this level.

Level 2: Match semantically compatible attributes. Some independently developed information systems use the same terms for the same concepts, but many don't. Sometimes, these differences in meaning are quite subtle. For example, in one system, "respondents at scene" may include only local respondents, including volunteer (untrained) civilians, whereas in another system, it may include all local, state, and federal respondents, but only those "officially" designated (no untrained civilians). If users combine results across systems without understanding these details, the resulting data is unlikely to satisfy the needs of the application.

Today, most matching is done manually in "data crosswalks," with the results typically captured in a spreadsheet. Large-scale data crosswalks can be enormously time-consuming, sometimes taking many staff months. In addition, the expensively gathered knowledge is not treated as an enterprise resource; as a result, future interoperability efforts that involve some of the same systems have difficulty leveraging prior match knowledge.

Recently, semi-automated schema matchers [11, 12, 17, 20] have been developed to reduce the time required to address Level 2 challenges. However, most of these have been either stand alone tools or they exist in a single-vendor stovepipe. To get the most leverage from a schema matcher, it should be integrated with an enterprise metadata repository and should interoperate easily with downstream tools that address heterogeneity at Level 3. As described in Section 2, this is a key aspect of our work.

Level 3: Map across diverse representations. Once semantic correspondences have been established, integrators still must reconcile different representations of the same concept. For example, For example, in international response, one team's system might measure distances in metric units (e.g., meters, kilometers) while their neighboring country measures in U.S. units (e.g., feet, miles). In addition, restructuring is often required, such as converting a deeply nested XML message to a normalized relational database schema. Finally, there are often many possible ways to combine related information (such as union, inner join, outer join). It currently requires considerable time from skilled programmers in consultation with subject matter experts to do this restructuring and combination in a way that meets consumer needs.

Like with schema matching, fully automated mapping is not possible, however, recent tools [6, 14] provide considerable help to integrators. Unfortunately, these tools typically lack integration with enterprise repositories and also exist in single vendor stovepipes.

1.2. Information Provenance

As more data is made widely available, consumers are able to gather vast quantities of information from many sources. Users must understand the information's provenance (i.e., where the information came from and the processes that acted upon that data) to determine if the information is useful and trustworthy. Additionally, information providers may augment the provenance information with additional metadata to help consumers interpret the information correctly.

Current provenance research typically attacks the problem from one of two perspectives: *database management* or *workflow*. Database provenance work [1, 2, 4] has traditionally focused on data-driven business processes and relies on the ability to trace information flows through SQL manipulations. Workflow provenance research [5, 9, 13, 19], on the other hand, generally tackles scientific processing and deals with less transparent, pre-defined process executions. The diversity among organizations with roles in emergency preparedness means that neither existing research thread by itself is sufficient.

In particular, provenance services for emergency preparedness must address the following additional requirements [3]:

- *Heterogeneity*: the approach must accommodate relational databases, XML, and monolithic files. In addition, a single data manager cannot be assumed.
- *Bi-directional provenance traversal*: it is important to reason about provenance in both the backward ("how was this data derived?") and forward ("which data depends on this?") directions.
- *Variable granularity*: different component systems will manage data objects and provenance at different levels of granularity (e.g., tuples, tables, or whole databases for relational data, and arbitrary size XML subtrees).
- *Incomplete disclosure*: Systems must be able to sometimes restrict views of provenance information, due to either privacy or security.
- *Confidence and accuracy*: Support is needed for originator estimates of accuracy and the possibility of alternatives; the ability to include additional annotations by subsequent users is also desirable.

1.3. Application to EP&R

Situation awareness is critical to effective emergency response, and both information interoperability and the improved information understanding enabled by provenance are critical to situation awareness. A common challenge among responders at an emergency is information sharing between two adjacent regions of operations (overseen by different regional emergency responders) or between different essential service functions. In the case of the former, a common (but not sole) issue arises in the use of a wide variety of incident commands systems. WebEOC (http://www.esi911.com) and ETeam (http://www.nc4.us/ETeam.php) are two popular examples of this type of software. Since each locality has flexibility to purchase, deploy, and train on the ICS they find best for their needs (including familiarity, ease of use, and financial considerations), they frequently find themselves limited in their ability to easily and dynamically share information with their colleagues in the neighboring region using a different system. While the situation is improving, currently these ICSs use different data formats, (security) access protocols, and varying communication standards. A particular challenge has been the exchange of geographic information (GIS) data. Multiple organizations may not share the same map sources, or may share them at different resolutions. Combined with limits in system interoperability to transmit and ingest the information appropriately, peer organizations typically resort to sharing static JPG or PDF maps and images which cannot be loaded, integrated, and modified between sources. This often results in each command center —holding up" 2 (or more) physical maps: their own, and one for each of the other responding command posts. As recently as last year, at a planned event in New England, state and local authorities in multiple command posts, *using the same major software*, could not easily share SA information due to the way the software was architected.

Of course, even when this information is shared, additional factors, some subtle, can limit the overall effectiveness. An effective provenance program brings critical trust and effectiveness to a rapidly evolving situation. Consider the previous challenge with sharing geographic information and asset location. When asset deployment information is shared on differing maps, it creates confusion and doubt in decision making, not just at the moment of sharing, but ongoing. What should a commander do when a shared asset map shows a —new" road where his own map does not? Rely on the presence of this potential shorter route? Assume that the more recent information is correct? What if it is simply mistaken? Rerouting may cost precious time. When an organization has to base decisions on information, such as maps, of unknown origin and reliability, it affects their confidence and ability to best allocate their resources.

While an actual emergency is the true test of information sharing, much key work must be done prior to a crisis. Increasingly, federal, state, local, and tribal authorities are working to improve preparedness. The players vary from region to region (e.g., the Coast Guard would be significant players along the New England coast, but not in Oklahoma City). The time for laying the information sharing foundation is during emergency preparation, training, and exercises. Removed from real time constraints and consequences, emergency organizations should take the time to document and share the systems they use, the critical data elements, schemas, and metadata (discussed further in Section 3). When exchanging all such information, in detail, across all partners is unrealistic, exercises provide the opportunity to identify the most critical sharing needs.

2. MITRE RESEARCH

We now describe research efforts at The MITRE Corporation that address the challenges described above.

2.1. Information Interoperability

Information interoperability (II) has long been a challenge for our government customers. Examples include the need for better information sharing among the Army, Navy, Air Force, and their coalition partners; the IRS seeking to consolidate and modernize its many legacy information systems; or the Federal Aviation Administration wanting to better share safety and airspace management information with international aviation authorities. Over the years, MITRE has often partnered with our customers to improve II. In addition, it is a priority area for research investment.

Our recent II research began with development of the Harmony schema matcher [15, 17]. Like all schema matchers, Harmony addresses Level 2 interoperability challenges by suggesting semantic correspondences across independently developed systems. An integration engineer then examines these suggestions, edits them as needed, and makes them available to downstream processes, such as creating code to perform data exchange. Novel aspects of Harmony include:

- The use of linguistic techniques to better exploit text documentation. Most prior work assumed that documentation would be missing (or erroneous), but we have generally found this not to be the case among our government customers.
- A user interface that allows the integration engineer to focus her attention on part of the

matching problem. Using Harmony, she might first identify high-level correspondences and then delve into the details of a specific subschema.

While working with Harmony, we discovered that integrating Harmony with other integration tools (e.g., those that assist with Level 3 challenges) was too hard. The II community lacked standards that would allow II tools to interoperate. In response, we proposed the Harmony Integration Workbench and collaborated with BEA to use the Integration Workbench to integrate Harmony with BEA's AquaLogic Data Services Platform (ALDSP), a commercial mapping tool [7]. This work demonstrated the viability of the workbench approach, allowing tools to cooperate by exchanging knowledge through a metadata repository rather than point-to-point tool interfaces.

In parallel, we began to explore the role standard schemas play in II and observed a common design pattern: Many successful standards begin with a small core data model that applies to a wide range of activities. This core is then extended in different directions by various sub-communities with more specialized interests. We refer to this design pattern as *core + corona* [18].

A key advantage of the core + corona pattern is that, in the core, you get II for free. Unfortunately, developers are accustomed to starting from a blank design slate. The Galaxy tool[16] provides developers with a number of features to encourage reuse using the core + corona pattern. First, developers use Galaxy to search a metadata repository for existing data models that partially meet their information needs. Second, they use Galaxy to customize the chosen model. Finally, Galaxy can automatically provide Level 3 interoperability over the shared portions of the model.

Often, an appropriate core + corona data model does not already exist for a community. Currently, there are few tools to speed the development of such models. In response, we developed the Common Ground Workbench (CGW) to support the following workflow:

- CGW ingests community schemata into a metadata repository.
- Integration engineers use the Affinity clustering tool to identify groups of schemas likely to share common concepts (e.g., "medical record", "passenger list").
- For each group, integration engineers use Harmony to determine semantic correspondences across that group (i.e., concepts common to all schemas, concepts common to all but one, etc.).
- CGW then exports a core + corona data model for the group, based on these correspondences.

The exported model can be used to establish a community exchange schema. In resource-constrained circumstances (e.g., where speed is vital), data exchanges can be based on only the core; an exhaustive integration is not necessary.

Testing the Common Ground Workbench on realistic Homeland Security schemas required us to extend Harmony to handle "industrial strength" schemas involving many thousands of data elements. Enhancing Harmony had the unexpected effect of enabling decision makers to plan large II tasks more effectively [22]. Many of the aforementioned tools require some sort of metadata repository to store knowledge about schemata and mappings. By providing these tools with a common repository, we can allow these and other II tools to interoperate more easily.

MITRE is now partnering with industry and academic leaders (including Google, the University of California at Irvine, and the University of Wisconsin) to develop OpenII—an Eclipse-based framework for II tools. At the heart of OpenII is the SchemaStore metadata repository based on the Harmony Integration Workbench. II tools can communicate with one another through SchemaStore, or more directly via the Eclipse framework. As a result, integration engineers are able to choose the best II tools for their specific task.

OpenII also includes importers for a variety of schema types including relational databases, XML Schema, and the Web Ontology Language. Within SchemaStore, schemas (and the mappings among them) are represented in a neutral extended entity-relationship model called M3 (MITRE Meta-Model). As a result, many of the tools work on a variety of schema types, and mapping information is also reusable across these different technologies.

OpenII is freely available as open-source software. Harmony, Galaxy and the Common Ground Workbench are all also freely available as components of the OpenII framework. Based on our collaboration with industry and academia, additional OpenII capabilities will be available from other sources. Moreover, because OpenII uses the Apache license, vendors can incorporate the framework into their products and provide a migration path to more industrial strength solutions (e.g., high-performance distributed query processing). However, the knowledge gained will still be accessible in vendor-neutral form in the metadata repository and can be shared across tools.

Two additional OpenII tools are under development addressing Level 3 challenges (i.e., mapping across diverse representations): RMap, which helps integration engineers create SQL mapping code, and XMap, which does the same thing using XQuery. MITRE is developing

RMap, while XMap is being developed by UC-Irvine with assistance from Google and MITRE. While mapping tools are at least partly model-dependent, we are attempting to maximize commonalities between RMap and XMap.

The initial release of OpenII will be August 2009. Visit http://openintegration.org for more information.

2.2. Provenance

As noted above, many of our customers are pursuing large-scale efforts to increase information sharing, whether for EP&R, counter terrorism, disease surveillance, law enforcement, or coalition military operations. A common theme is to increase the visibility, accessibility, and understandability of information, not just to known consumers but also to unanticipated beneficiaries of the information throughout a very large, multi-organizational mission space. An example is the Department of Defense's Net Centric Data Strategy (NCDS) [10].

With greater information access, though, comes a new challenge; with increasing numbers of sources outside of users' typical sources, they may have difficulty assessing source trustworthiness. An essential ingredient in making that trust determination is provenance—i.e., where the information came from. In recognition of this, the NCDS asks information providers to describe the derivation of all posted data resources, so that "the pedigree of each data asset is known and available."

Unfortunately, the current state of the practice is simply to provide a manually populated metadata field for provenance or pedigree. Not surprisingly, this information is rarely provided, because there are no practical tools to automate the collection of provenance information in heterogeneous, multi-organizational environments.

In recognition of this gap, MITRE initiated the PLUS project [3]. PLUS is developing a service that collects provenance information from participating systems, maintains the "family tree," and allows users to pose queries over the provenance information. The PLUS provenance service is being designed to collect lineage information with minimal development time for participating systems and without affecting normal operation of legacy systems. PLUS extends prior database and workflow provenance research to include web information and complex processes, such as those used in data fusion applications.

Through the provenance family tree, a resource (data or process) is associated with other ancestor or descendant resources, and these resources may be subject to different security and privacy release policies. One goal of the PLUS project is to provide as much provenance information as possible while enforcing these policies. To that end, PLUS allows resource administrators to define more widely releasable *surrogates* that contain alternative information about a resource. Examples of surrogates are subsets of a relational table, a general description of a process rather than a detailed algorithm, a redaction of a document, or even a *signpost* that gives contact information for negotiating access. PLUS also allows limited surrogates for the provenance relations among resources; these summarize relationships without giving details of intermediate resources. Using the family tree analogy, such a surrogate edge might link a grandparent directly with a child, omitting information about the parent between them. This technique retains the connection to other ancestors (e.g., the great-grandparents) while enforcing release policies about the intervening resources.

Provenance information helps consumers understand and trust data, but it is also enables a variety of useful analysis tools. In the face of corrupt or inaccurate data (whether due to malicious attack or error), PLUS allows authorized users to propagate warnings to downstream process and derived data resources. We are also exploring the use of provenance information to reconstitute corrupted data (essentially replaying previous operations) and the possibility of providing surrogates containing out-of-date or approximate versions in place of corrupted data.

Provenance information is also a useful record of which resources are used most often or support critical mission activities and downstream data assets. System analysts can use this information to identify resources that should be made highly available or require greater protection against cyber attack. Additionally, the provenance family tree can aid discovery, by giving users information about related resources, e.g., other data derived from some of the same ancestors.

3. APPLYING THE EMERGING CAPABILITIES

This section describes how the emerging capabilities just described could enhance information sharing and understanding in an emergency. We describe how the various pieces could be employed during preparation, training/exercises, ongoing operations, and emergency response.

3.1. Establish Infrastructure

To get the benefits of information interoperability and provenance tools, it is helpful to establish some

infrastructure in advance of an actual crisis situation. Specifically, EP&R stakeholders (e.g., emergency operations centers or regional fusion centers) would stand up two pieces of software: (1) a metadata repository containing the schemas of participants likely to share data in an emergency, and (2) a provenance store for capturing provenance family trees. We now describe how this would play out with the OpenII, Common Ground Workbench, and PLUS tools.

Although all participants may not be known yet, establishing a repository and populating it with an initial set of schemas jumpstarts the response to an emergency. Because SchemaStore is platform independent, and its importers access many types of schemas, the major tasks are: a) to identify likely participants (already part of the preparation process) and b) to obtain their data models.

Once a repository is established, two questions must be answered: "What are likely sharing communities?" and "What are they likely to share?" These are both addressed by the Common Ground Workbench.

Because sharing partners typically have overlapping schema contents, clustering schemas based on those overlaps can help identify partners. The Affinity component of CGW clusters schemas in the repository; users visualize these clusters and manually adjust them until each cluster accurately reflects a community of sharing partners.

Harmony is then used to determine a simple core model for each community. Our strategy in the preparation phase is to focus on the core model (concepts that most partners felt were important enough to include) because a) the core is likely to be stable over time and b) it establishes basic interoperability with minimal effort (i.e., the rewards of sharing are not obscured by a seemingly interminable process).

Finally, using OpenII's XMap and RMap tools for creating executable data transfer mappings, code is generated allowing each partner to produce and consume data formatted according to the core model, enabling necessary data exchanges during a crisis.

To get the benefits of provenance, the first step is to stand up a provenance store such as that developed under the PLUS project. Second, administrators must determine appropriate provenance capture points. Clearly, one does not want to have to modify legacy systems to do this; instead, it is important to identify more general capture points, such as an Enterprise Service Bus (ESB) or a Business Process Execution Language (BPEL) engine. This has been the strategy we have used on the PLUS project, in which we demonstrated lineage capture from legacy systems by providing hooks into the MULE ESB.

When some of the provenance information is sensitive, the preparation phase is when stakeholders specify provenance release policies [21]. For example, pharmacy chain P may be willing to provide certain information to a bio-threat surveillance system under the condition that its data contributions are known only to cleared participants. In that case, administrators could create a surrogate for resulting provenance records, indicating merely that the downstream products came from "pharmacy data" without identifying the specific chain.

3.2. Training/Exercises

Training exercises that simulate response scenarios are essential to emergency preparedness. From the perspective of II, these exercises stress a) the ability to locate sharing partners with information relevant to the scenario, b) the ability to share information among these partners and c) the ability to reason about provenance: which information was provided by the various partners.

We assume that during the previous phase, the various EP&R organizations have had the opportunity to populate their metadata repository. If not, the first II task is to complete the first phase as quickly as possible using OpenII/Common Ground tools.

Given the details of the training scenario, the response team needs to determine which sharing partners need to coordinate their efforts. To some extent, this decision depends on geographic constraints such as where the simulated event transpired. Other constraints depend on the nature of the scenario. For example, in the event of pandemic flu, the responders are not interested in information about nuclear reactors, but in the event of a natural disaster (e.g., an earthquake), such information might be highly relevant.

The OpenII/Common Ground tools provide search [8] and clustering capabilities over the metadata repository; these capabilities allow the responders to identify relevant sharing partners and the related interchange schemata. An ongoing research challenge involves determining how much of a schema to return based on a given search: if entire schemata are returned (as is currently the case in OpenII), the responders may not be able to easily determine how the result is relevant to the exercise. If individual data attributes are returned, the responders may not have enough context to determine how to use the search results.

Once sharing partners have been identified, the next task is to create (and deploy) mappings that will allow information exchanges to transpire. In the previous phase core-based interchange standards were developed, and key data assets were mapped to the interchange standards.

As a result, with no further effort, some level of interoperability is immediately available. Additional interoperability can be obtained at this point by expanding this process to support exchanges beyond the core (i.e., among members of a subset of the sharing community) that the exercises identified as of especially high value. (Because these mappings are stored back into the metadata repository, they are available should these sharing partners need to respond to an actual emergency.)

Sharing partners can begin reporting provenance information to the PLUS provenance service whenever they are ready. The earlier they begin, the richer the provenance family tree will be, but even a little provenance information can provide useful context. Similarly, as new partners are identified, their provenance information enhances the extended family tree stored by PLUS. The provenance information contributed by sharing partners reflects the growing use of shared resources and provides insight into sharing activity.

The ability to analyze information flows and sharing is particularly important in exercises; provenance is a key instrumentation that plays a critical role in after-action reviews. Together with other observations and instrumentation, collected provenance information can help identify information gaps and sharing shortcomings. Provenance information can be matched against participants' reports to identify areas for further analysis. Based on these reviews, provenance can contribute to improved situational awareness, revised information sharing, and increased availability of key resources.

3.3. Ongoing Operations

Once the basic infrastructure is established, one can begin using these capabilities in normal, day-to-day operations. Over time, sharing partners come and go and schemas evolve, requiring occasional revisions to the core schemas and mappings. In addition, periodic reviews will establish new interoperability opportunities, such as new high value sharing partners.

In some cases, requirements will emerge for *composite data services* that provide a fused view of multiple sources. Composite data services require not only interschema mappings but also policies for merging and deconflicting information. For example, if systems A and B report FloodDamage as "minimal" and "severe" respectively, what should be shown to the user? Composite data services are typically implemented as a backend data warehouse that is made accessible via a web service or portal. The warehouse schema would typically conform to (or be a superset of) a core schema developed in the previous phase, and the warehouse would be populated based on the mappings to that core schema.

Provenance collection that was established during the first phase (establish infrastructure) would continue during ongoing operations. This should be reviewed periodically to see if provenance is being captured at the right granularity. If it is captured too coarsely, it may not provide enough detail to give users adequate context; it may also be insufficient for understanding the impact of (and recovering from) corrupted data. On the other hand, excessively fine grained provenance collection could overwhelm users with uninteresting context (for example about trivial message reformatting) and result in unnecessarily large storage costs.

3.4. Response

In the event of an actual emergency, the same tasks identified for training need to transpire. Relevant sharing partners need to be identified and II must be established among those partners. The speed with which II can be established depends on how much work was completed in earlier phases. Nevertheless, actual emergencies rarely play out entirely according to plan. Having already established a metadata repository, a core interchange schema, and some mappings provides participants tools for quickly establishing additional high-value II.

When possible, provenance information should be collected to facilitate after action review. However, if the overhead of collecting provenance information is too great, this component should be turned off to allow timely response to the emergency. Realistic provenance information can be collected after the fact by creating a training exercise that mimics the actual event.

4. CONCLUSIONS

This paper has described two important information sharing challenges for EP&R: information interoperability and provenance. We presented MITRE applied research that is addressing key gaps. We also illustrated how the products of this research support improved information sharing and understanding through the different phases of EP&R.

While some aspects of our work are on the "bleeding edge," we also have produced fairly mature capabilities that are ready for application. In particular, the OpenII and Common Ground Workbench tools have schema match and clustering capabilities that have already been applied to customer problems. In addition, the first full release of the OpenII toolkit is scheduled for August 2009. Future releases will include data exchange code generation for both SQL and XQuery. The PLUS system is being tested in a large command and control simulation environment, and we are also planning a pilot in a biosecurity monitoring system. We actively seek

collaborations to help us refine our tools and apply them to improve information sharing for a wide range of homeland security missions.

ACKNOWLEDGMENTS

Thanks to Chris Wolf, Adriane Chapman, Arnie Rosenthal, M. David Allen, Doug Burdick, and Michael Morse for many helpful discussions and implementation of these concepts. In addition, thanks to our OpenII collaborators, especially Alon Halevy, Mike Carey, Jayant Madhavan, and AnHai Doan.

REFERENCES

[1] O. Benjelloun, A. D. Sarma, A. Halevy, and J. Widom, "ULDBs: Databases with Uncertainty and Lineage," *VLDB Seoul, Korea*, pp. 953-964, 2006.

[2] D. Bhagwat, L. Chiticariu, W.-C. Tan, and G. Vijayvargiya, "An Annotation Management System for Relational Databases.," *VLDB*, pp. 900-911, 2004.

[3] B. T. Blaustein, L. Seligman, M. Morse, M. D. Allen, and A. Rosenthal, "PLUS: Synthesizing privacy, lineage, uncertainty and security," *ICDE Workshops*, pp. 242-245, 2008.

[4] P. Buneman, S. Khanna, and W.-C. Tan, "Why and Where: A Characterization of Data Provenance," *ICDT*, pp. 316-330, 2001.

[5] S. P. Callahan, J. Freire, E. Santos, C. E. Scheidegger, and C. T. S. H. T. Vo, "VisTrails: Visualization meets Data Management," *SIGMOD*, pp. 745-747, 2006.

[6] M. J. Carey, "Data delivery in a service-oriented world: the BEA aquaLogic data services platform," in *ACM SIGMOD*, 2006.

[7] M. J. Carey, S. Ghandeharizadeh, K. Mehta, P. Mork, L. J. Seligman, and S. Thatte, "AL$MONY: Exploring Semantically-Assisted Matching in an XQuery-Based Data Mapping Tool," SDSI, Vienna, Austria, 2007.

[8] K. Chen, J. Madhavan, and A. Halevy, "Exploring Schema Repositories with Schemr," *ACM SIGMOD*, 2009.

[9] S. Davidson, S. Cohen-Boulakia, A. Eyal, B. Ludascher, T. McPhillips, S. Bowers, and J. Freire, "Provenance in Scientific Workflow Systems," *IEEE Data Engineering Bulletin*, vol. 32, pp. 44-50, 2007.

[10] Department of Defense, "Department of Defense Net-Centric Data Strategy," May 2003.

[11] H. H. Do and E. Rahm, "COMA - A System for Flexible Combination of Schema Matching Approaches," VLDB, Hong Kong, China, 2002.

[12] A. Doan, P. Domingos, and A. Y. Halevy, "Reconciling Schemas of Disparate Data Sources: A Machine-Learning Approach," *SIGMOD*, 2001.

[13] P. Groth, S. Miles, and L. Moreau, "PReServ: Provenance Recording for Services," *Proceedings of the UK OST e-Science second All Hands Meeting 2005 (AHM'05)*, 2005.

[14] M. Hernandez, R. J. Miller, and L. M. Haas, "Clio: A Semi-Automatic Tool For Schema Mapping," *SIGMOD Record*, vol. 30, pp. 78-83, 2003.

[15] P. Mork, A. Rosenthal, L. J. Seligman, J. Korb, and K. Samuel, "Integration Workbench: Integrating Schema Integration Tools," InterDB, Atlanta, GA, 2006.

[16] P. Mork, L. J. Seligman, M. Morse, A. Rosenthal, C. Wolf, J. Hoyt, and K. Smith, "Galaxy: Encouraging Data Sharing Among Sources with Schema Variants," in *International Conference on Data Engineering (ICDE) 2009*. Shanghai, China, 2009.

[17] P. Mork, L. J. Seligman, A. Rosenthal, J. Korb, and C. Wolf, "The Harmony Integration Workbench," *Journal on Data Semantics*, vol. 11, pp. 65–93, 2008.

[18] P. Mork, J. Stanford, L. Seligman, J. Hoyt, and K. Smith, "Jump-Starting Data Integration in Biomedicine," *Workshop on Data Integration in the Life Sciences (DILS 2007)*, 2007.

[19] T. Oinn, M. Greenwood, M. Addis, M. N. Alpdemir, J. Ferris, K. Glover, C. Goble, A. Goderis, D. Hull, D. Marvin, P. Li, P. Lord, M. R. Pocock, M. Senger, R. Stevens, A. Wipat, and C. Wroe, "Taverna: lessons in creating a workflow environment for the life sciences: Research Articles," *Concurr. Comput. : Pract. Exper.*, vol. 18, pp. 1067-1100, 2006.

[20] E. Rahm and P. A. Bernstein, "A Survey of Approaches to Automatic Schema Matching," *The VDLB Journal*, vol. 10, pp. 334–350, 2001.

[21] A. Rosenthal, L. Seligman, A. Chapman, and B. Blaustein, "Scalable Access Controls for Lineage," *First Workshop on Theory and Practice of Provenance Systems (TaPP)*, 2009.

[22] K. Smith, M. Morse, P. Mork, M. Li, A. Rosenthal, D. Allen, and L. J. Seligman, "The Role of Schema Matching in Large Enterprises," CIDR 2009, Asilomar, CA, 2009.