Layered Information Correlation and Aggregation Architecture

Peter A. Smyton
The MITRE Corporation
Mail Stop S221
202 Burlington Road
Bedford, MA 01730
(781) 271-7992
psmyton@mitre.org

Abstract

With the increasing proliferation of information and the ability for its broad dissemination comes the challenge of managing diverse information and providing appropriate aggregated subsets to individual users to allow and aid purposeful action. This paper describes an information correlation and aggregation architectural structure for where and how information should be assembled into useful products. The paper addresses the benefits of such an approach, such as flexibility, ease of expansion, improved interoperability and mission execution, and efficiency in use of resources. The technologies, both current and future, that need to be applied to the architecture are identified (e.g., data mining, personalized subscription services, plug-and-play infrastructure to support real-time information management, information combining techniques, etc.). The architecture within the information supplier and consumer nodes is described, and a suggested functional decomposition of the information management domain into components is proposed. Finally, key issues that need to be addressed, such as pedigree, the role of advantaged and disadvantaged consumers, and data looping, are discussed.

1. **Problem Statement**

The development and continued growth of the Internet has had a profound impact on the access and dissemination of information. Leveraging off of this technological breakthrough, much of the DOD has embraced the concept of a global grid, where each user can access incredible amounts of information simply by plugging their information appliance into the network. Further, other technological advances are becoming operational in the military, such as increases in wide coverage space-based detection and information gathering capabilities, proliferation of UAVs, with multiple sensor suites, the use of multilevel security to allow dissemination of sanitized highly classified data for tactical purposes, etc. These advances are resulting in an increasingly diverse data rich (though some would claim data saturation) environment. With this increasing proliferation of information and the ability for its broad dissemination comes the challenge of managing the information and providing appropriate subsets of the information to individual users to allow and aid purposeful action.

Clearly there are as many purposes for the information as there are missions. For example, the Joint Theater Air and Missile Defense Organization (JTAMDO) architecture categorizes air and missile defense systems into three broad classes, each with its own data needs and timelines:

- Weapons/sensor systems, which include fire control within its missions, require plot level sensor detections with data latencies on the order of one second

- Command and control (C2) systems, which include situational awareness within its missions, tend to need a combination of plot level and track level information with latencies on the order of seconds or tens of seconds

- Theater systems, which include planning within its missions, need a broad view of aggregated data across an entire theater, but can usually tolerate latencies on the order of minutes or greater

An issue that the services currently struggle with, and will become more problematic in the future, is how to structure the management of information, so the right information is available to the right user in the right form at the right time. More specifically, where and how will information from diverse sources be correlated and aggregated and by whom, given the wide variety of uses.

2. **Relevance to Command and Control**

The rapidly changing world of command and control has recently embraced the concept of a single integrated C2 system (IC2S). The concept of the IC2S, from Joint Vision 2010, is to construct specific systems from common components and capabilities, thereby reducing the inefficiency associated with development of stovepipe systems with redundant capabilities, and enhancing interoperability. A 1998 USAF Scientific Advisory Board (SAB) Summer Study on Information Management to Support the Warrior underscored the importance of information management to the IC2S and C2 in general [McCarthy, *et al.*, 1998]. It offered a publish-and-subscribe model as the major construct for information management, but did not address the details of how diverse information would be combined. An architecture that promotes full interoperability and efficient information management is critical to the successful prosecution of the IC2S mission. This architecture must satisfy a number of tenets important to C2. First of all, the architecture must be sufficiently flexible to perform information correlation and aggregation tailored for different missions and a different mix of C2 platform components, including missions and component mixes that one cannot currently identify or predict. The architecture needs to be adaptive to easily include new information sources, and information processing applications and management. The architecture should be designed to allow the use of common tools across multiple platforms, so that consistent information products and views are possible. Finally, the architecture should allow for the seamless transition between phases (e.g., planning and execution), with latencies appropriate for the task, so that the information and tools that are relevant for multiple phases of the mission can be used without confusion or interruption.

3. **Layered Architecture**

Using the tenets described above and an understanding of information management concepts employed and emerging today, an architectural structure for where and how information should be created and managed in the future, with an emphasis on the correlation and aggregation of information, has been developed. This structure uses a layered concept to understand the context in which information products of the future must reside, and is consistent with the SAB study. The architecture consists of three layers.

3.1 ***Producer Layer***

In layer 1, systems that generate information, called producers, collect information from their organic sources, and perhaps correlate multiple sources and combine the information. This fusion would be done "locally" for a number of reasons:

- For the local system's use (e.g., sensor control and optimization)

- To eliminate redundant, unimportant, or stale data that would not be generally useful in the Battlespace InfoSphere

- To provide aggregated information products to disadvantaged users that connect directly with that producer

A disadvantaged user is one that may not be able to process the "raw" data directly, due perhaps to inadequate processing capabilities or the need for aggregated information quicker than they can develop themselves (e.g., time critical targeting).

The information potentially relevant to others is pushed out into the Battlespace InfoSphere (BI). The BI is the virtual single, distributed repository containing all of the information needed to perform the scope of missions across the battlespace. This push should allow for the possibility of both the "raw" source information (e.g., radar plots) and processed data to be deposited in the BI. There are different types of ultimate, downstream users of the information, called consumers, some wanting raw data (advantaged users) and some only needing the processed data (disadvantaged users). In both cases a minimum suite of information should be included:

- Location (geo-spatial) and time (temporal). To maintain a common structure for all consumers, it is important that these two types of information be included with all information. This information can serve as the underlying basis for registration and correlation, and ensuring everyone has a common set of reference data when sharing information.

- Component sources used to produce the information and a series of data quality indicators (e.g., accuracy, perishability, use of the information, etc.). The specific data quality indicators will depend on the types of uses of that information by the consumers.

- Attributes of the information (e.g., characteristics, like amplitude, shape, identity, etc.).

A notional illustration of this layer of the architecture is shown in Figure 1. As can readily be seen, this architecture is distributed. Though there are advantages to a centralized approach (e.g., single set of functions do the correlation, resulting in common products), the distributed approach is consistent with the current information architecture. For example, use of local data

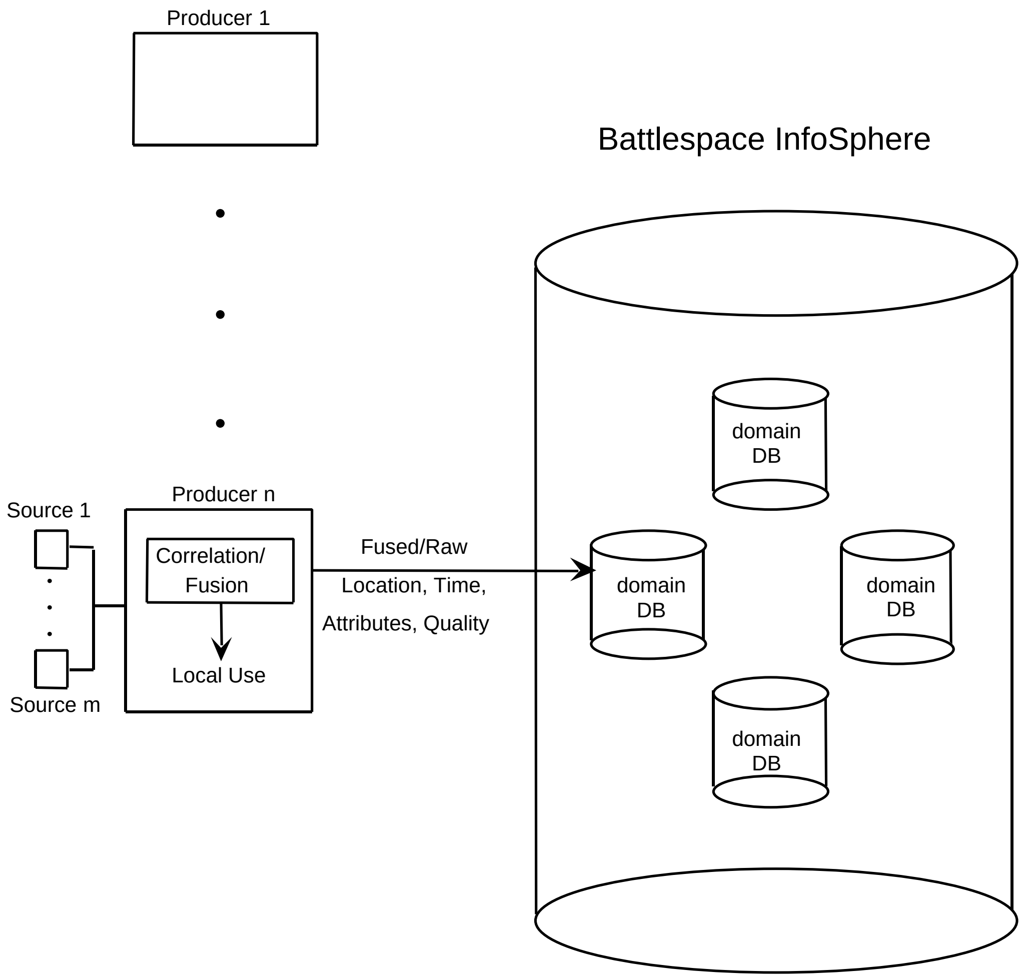

Figure 1. The Producer Layer

for correlation and fusion in parallel with depositing the data in the BI is consistent with most link protocols (e.g., Link-16). The use of local data also implies lower latencies, perhaps required as part of the producer's mission. Further, recognizing that a correlator's best performance is achieved when the system is tuned to the information sources, it is logical for the initial correlation process to be done on the producer's platform, as close to the originating sources as possible. However, since raw data can be accessed by later users, a central correlation facility can be realized, and even take advantage of preliminary correlation done at the producer level.

3.2 ***Battlespace InfoSphere Layer***

In layer 2, the pushed information is placed in the BI, which can be thought of as a data warehouse. In reality, the BI would consist of a set of domain databases; one can think of these databases as being as diverse as a relatively static intel database, or a Link 16/TADIL J network, whose data is being dynamically updated. This BI has the advantage of being the virtual location where a broad set of information, gathered from diverse sources, all reside. A notional view of this layer is in Figure 2. It is at this layer that information collected from different producers

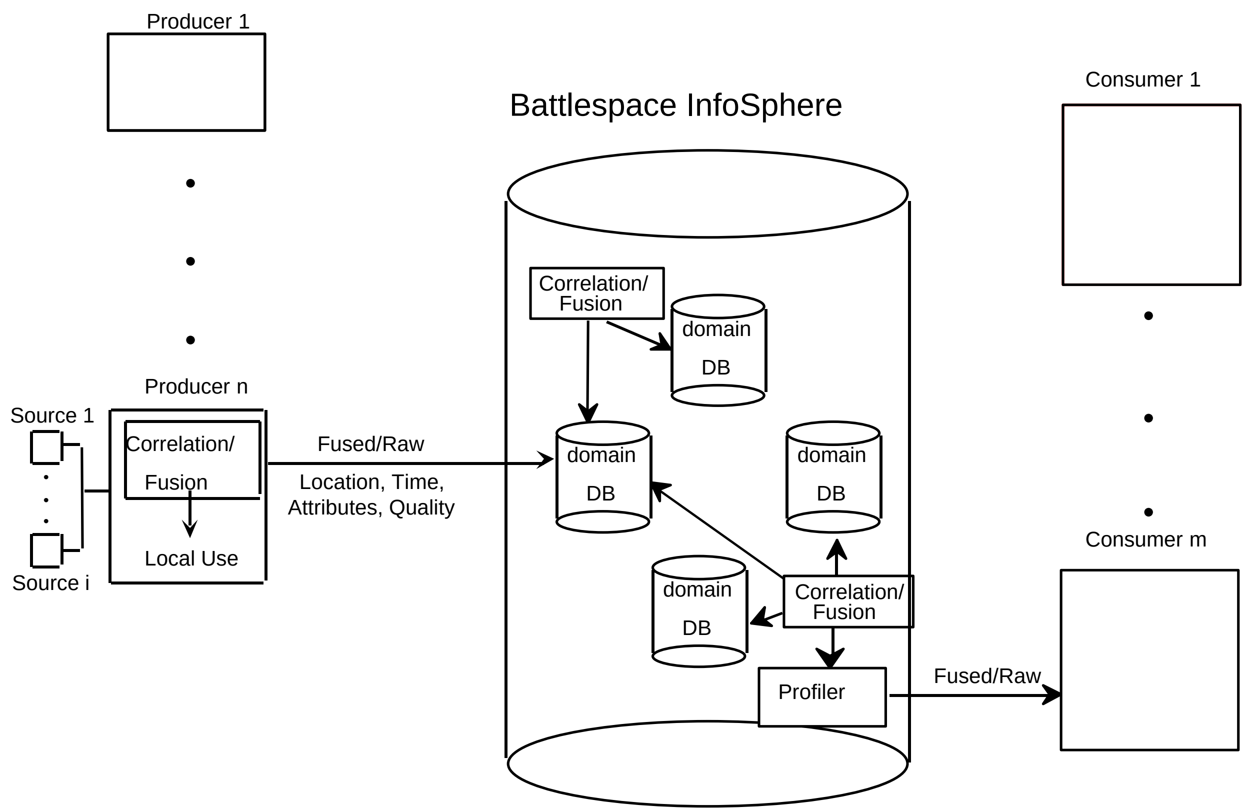

Figure 2. The Battlespace InfoSphere Layer

about the same object, or events that occurred at the same time/location, can be correlated and, to the extent possible, aggregated. The basis for aggregation may be the time of occurrence, for example. A current area of promising research in correlation involves the application of hybrid techniques [Kandel, 1992] [Doyle, 1996]. Acknowledging that widely diverse information types cannot be easily correlated using a single technique, the hybrid approach uses a series of different techniques against a given problem, each technique appropriate for the information set it is to process. The types of correlators and other information management tools that "reside" in the BI depend on the information product needs of the consumers. Thus, among the tools needed would be one, perhaps for each consumer, that would understand that consumer's specific product needs and would orchestrate or manage the collection and combining of the relevant information; we'll call that tool a profiler. The resultant "tailored" products would then be pushed to that consumer.

Components of the BI relevant to information management include:

- Domain databases. Examples includes the Modernized Intelligence Database (MIDB), the Navy's Cooperative Engagement Capability (CEC) network, the Link-16 network, the Air

Tasking Order (ATO), the Target Nomination List and the Rules of Engagement (ROEs). These databases can be characterized as any information repository that needs to be accessed by potentially multiple producers and consumers to accomplish the mission.

- Database maintainers. This software component would either be part of or adjunct to each database. Its function would be to maintain the quality of the information in the database by discarding incorrect, redundant, or perished data, updating information (e.g., position updates), and ensuring information is deposited correctly (e.g., new ID information is placed in the appropriate track file).

- Profilers. In an environment in which each consumer has specific, somewhat unique missions and needs, it is important to provide the right information at the right time to each consumer. To do so while moving towards common functional components is a challenge. A profiler can bridge that gap by understanding the unique information needs (content, timeliness) of the consumer, and selecting and tailoring products from the information within the BI (or tasking the agent described below to collect the information).

- Agents. Agents are tasked by each profiler to search through the BI database to find the information relevant to that profiler's consumer. An example might be to find all events that have occurred within a particular geographic area during a specified time period, or to find all information within a region that is pertinent to a class of mission (e.g., active missile defense).

- Publish/Subscribe capabilities. Publish/subscribe capabilities would work in conjunction with producers and correlators to post information in the BI, and with profilers to provide that information to the consumers who have a regular or on-going need for that information.

- Correlators and aggregation systems. These functions would attempt to correlate diverse information provided to it by agents which may concern the same object or event. Further, as appropriate, the function would combine correlated information to create "new" aggregated information that is directly applicable to one user's decision directed needs.

- Information brokers. In this architecture, much of the information is pushed directly to the consumer for their review and use. Information not normally needed or used is pulled by the consumer via information requests. These requests would flow through the profiler (who would determine if the profile should change to include the requested data be pushed in the future) and pass to the broker. The broker would then reconcile the best way to provide the requested information to the consumer.

The correlation and fusion done within the BI provides many of the benefits of a centralized architecture. Common functions/tools are used and thus common aggregated information products are generated and made available to consumers. Further, standard types of information, in standard formats, are made available to the consumer community. Though there is standardization, the use of the profiler allows the pushed products to be tailored to the needs of that consumer (e.g., mission, area of responsibility, etc.).

3.3 ***Consumer Layer***

In layer 3, the information products would be received by the consumer, processed and, perhaps, correlated/combined further, using organic or unique information available to the consumer that has not been provided to the BI. This information could be local intelligence not readily available elsewhere in a timely fashion. This layer is depicted in Figure 3. Note that this information that is provided from the BI to the consumer may still be in raw form, if the consumer wishes to do the processing needed to create the desired products themselves. In addition to this pushed information, the consumer can request ("pull") information from the BI.

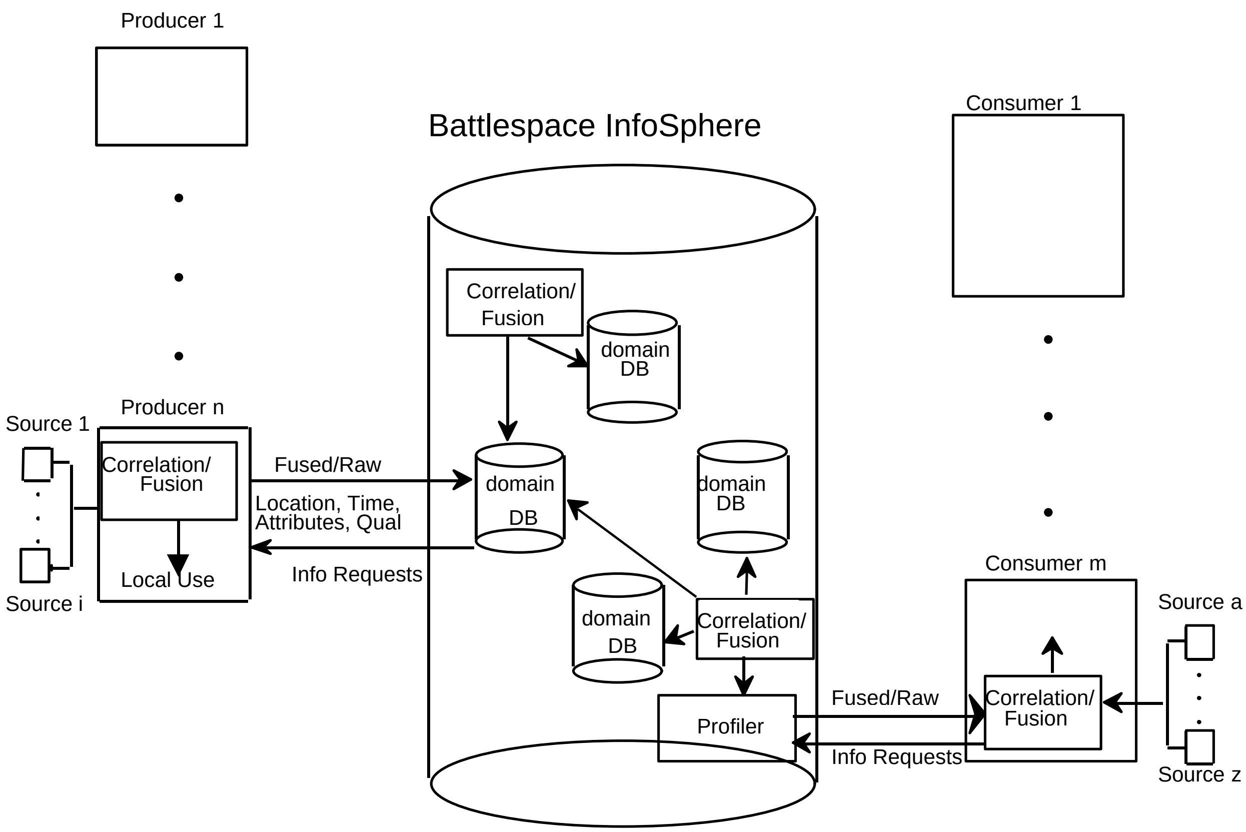

Figure 3. The Consumer Layer

This information would typically take one of two forms: (1) information that the profiler in the BI does not typically provide (e.g., amplifying data or data rarely needed by that consumer) or (2) requests for needed information that is unavailable immediately. In this latter case (e.g., a sensor tasking request), conceptually a broker (described briefly in the previous section), "residing" in the BI, would assess the request. The broker would then task one or more of the producers to supply the missing information (e.g., schedule the gathering of the data at an appropriate collection opportunity). The broker would decide which producer(s) to task based on a number of factors, to include the urgency of the request, that consumer's priority compared to others in the queue, the quality of the data needed, and the availability and capability of the producer to provide the needed information. An extension of this concept might be that when the nature of the request cannot be satisfied by a single information source, then the broker may task a correlator/aggregator system and multiple information producers to create a new "information product" which can satisfy the request.

Note that, since the BI is virtual, the objects (domain databases, tools) that reside in it can actually be located in systems within the Battlespace, and that a single system can assume multiple roles (e.g., producer and consumer) and perform multiple tasks (e.g., broker, profiler).

The capability to perform correlation and fusion at the consumer provides the same types of benefits as are provided by the distributed aspects of the architecture at the producer. Additionally, the use of the broker has the attributes of (1) allowing access to a common information set, (2) providing a standardized feedback mechanism for sensor tasking, and (3) facilitating system-wide conflict and error resolution.

3.4 *Architecture for an Example Application*

Consider how this layered architecture would apply in an application such as real-time Ground Moving Target Indicator (GMTI) based surveillance; note that the discussion can be expanded to address ground surveillance in general. Figure 4 shows a notional approach, with GMTI producers such as the Army's Airborne Reconnaissance Low (ARL) sensor, UAVs, space assets, and Joint STARS on the left, and consumers, such as the Air Operations Center, or AOC (to support Time Critical Targeting), a fire control center, and the ground view of the Common Operational Picture (COP) on the right. The COP is the CINC's view of the battlespace, drawing on all information collected within the BI. The COP can also be used to feed back corrections to the BI, producers, and other consumers. Sample domain databases are shown in the BI to include the ARL and Joint STARS datalink (SCDL) "report" or "plot" level databases, the Link-16 net that may provide GMTI-based tracks from several platforms, and an intel database (with known maintenance, fuel, supply, garrison locations, etc.). For readability, only some of the BI components are shown in the figure. One correlation/aggregation function could be to perform centralized tracking using all GMTI report level databases, that is, to form tracks from the plots contained in these databases. The Joint STARS producer is shown in more detail, using its organic source information (GMTI, Order of Battle, etc.) to form tracks for its local use. However, both plot data (via SCDL) and track data (via Link-16) could be disseminated to the BI, so that the type of information appropriate for a given consumer of GMTI data is available (e.g., raw report data for advantaged users and processed track data for disadvantaged users). Likewise, the ground view of the COP (perhaps at the CINC's Headquarters) is shown in more detail as a typical consumer. The information products are pushed to the COP by its profiler, and information requests or feedback (e.g., corrections to information based on insight available at the COP) are pulled, again going through the profiler.

This simple example of GMTI tracking could be broadened to include multiple source ground surveillance, including Combat Identification. In this broader example, diverse sources such as

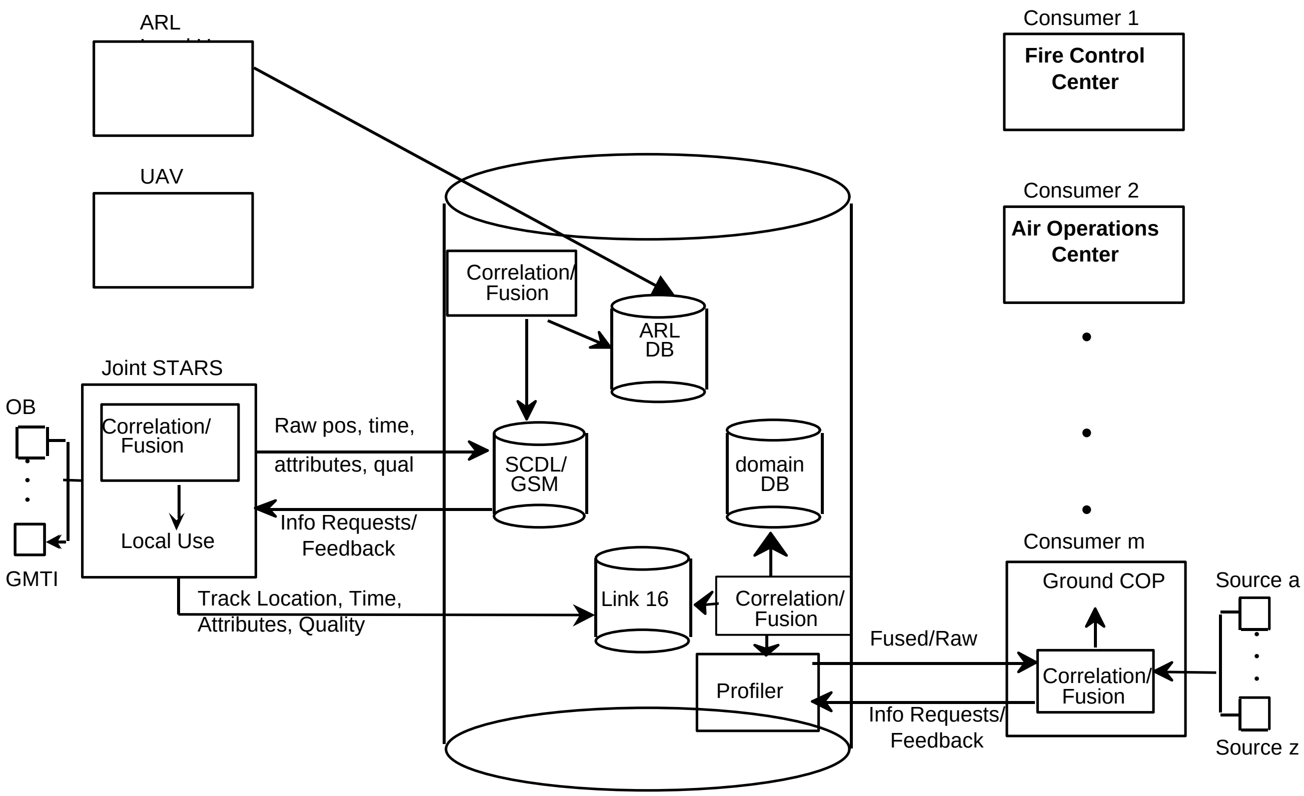

Figure 4. Use of Architecture in a GMTI Application

Synthetic Aperture Radar (SAR), stored intelligence data, ESM, EO/IR, or acoustic sources would be included, and play a key role in performing an expanded ground surveillance mission.

4. Future Challenges

4.1 *Technological Developments*

There are several technologies that need to be developed further for this architectural structure to become a reality. Though many of these technologies are already being addressed, further work needs to be done in basic research and in applied research before the technology finds its way into operational systems. The most critical of these are found within the BI, and include:

- Database maintainers: the capability to maintain the integrity of the database in a heavy load, dynamic environment.

- Software agents: applying data mining principles across the breadth of the distributed BI database.

- Profilers: providing robust personalized subscription services appropriate for a military environment. Work is on-going at MITRE in this critical area (information can be viewed at http://www.mitre.org/).

- Advanced correlators: though correlation and fusion are, relatively speaking, mature technologies, added sophistication is needed in some applications (such as hybrid approaches whereby multiple processing techniques are needed in the correlator to address the diversity of the information).

- Brokers: performing the complex task of forecasting optimal assignments in a multiple constraint environment.

- Plug-and-play infrastructure: possessing the underlying software that will allow the flexibility to add or replace components easily as the needs change, particularly in support of real time operations.

4.2 *Candidate Aggregation Components*

Under this architectural approach, information management functions common to multiple systems can be easily identified, and common application tools could be used, reducing the number of stove-piped systems with unique software components that perform similar or identical functions. One approach to cataloguing components is along functional lines. Relative to correlation and information aggregation, components could be categorized into: registration/gridlock, coordinate conversions, correlation, smoothing/filtering, ID estimation, situational awareness, threat assessment, and resource allocation. Within each of these components, a family of products may be needed to address the breadth of the needs within that functional area. In the case of correlation, this family could include airborne report-to-track, airborne track-to-track, missile report-to-track, launch point-to-missile track-to-impact point, etc., to name a few.

4.3 *Issues to be Addressed*

Even if technological advances and componentization of functions are realized such that the architecture proposed here is feasible, there are still several key issues that will need to be addressed before the goals of JV2010 are achieved. These include:

- Data pedigree. As information becomes more prevalent, users and systems will need better insight into the quality of the data. Currently, the number of quality attributes that are provided with the information is limited. One example is Track Quality in the Link-16 track message, but even that is merely a substitute for the covariance matrix, which would provide more information about the track's quality. Currently, pedigree is addressed in an ad hoc manner, mostly manually, based on what the user trusts (e.g., reputation) or assumes to be independent. The specific type of pedigree information desired is dependent on the application. Certainly information such as originating sources and their quality/accuracy, perishability of the information, and dependence on other information are key aspects of pedigree. To properly manage and use the huge amounts of information in the BI, knowledge of the pedigree is essential. Dissemination media (such as JCTN, JDN, etc.) need to address and account for pedigree in their evolution.

- The role and relationship of advantaged and disadvantaged consumers. Understanding the connectivity and relationship among systems within the IC2S, to include when a consumer is advantaged or disadvantaged, is important in information management. This knowledge is crucial for both the agent to know what information to access, and the profiler to know how to provide appropriate products to the consumer. Roles and relationships may be dynamic as well; having adaptive agents and profilers is important to ensuring the right information is delivered to the right consumers at the right time. Further, the extent to which the architecture can adapt, from mission-to-mission, day-to-day, and even dynamically in real time, as the mix of and relationships among consumer's changes, will need to be assessed as the architecture is adopted.

- Data looping. As network connectivity increases and increasingly more sources become available to a consumer, the risk of integrating the same or dependent information multiple times increases as well. Consider the simple case of a system receiving track data from a remote source via Link-16 and successfully correlating that track with a local track. If the remote and local tracks are fused, the resultant fused track will presumably be of better quality than either separate source track. At this point, the fused track would have a better track quality than the remote source track, and the local system will assume reporting responsibility (R2) and tell out the track. The remote source, having received the fused track and correlating it with its own track, could fuse the two tracks. Though the remote source believes the two tracks are independent, the remote source track state would actually be counted twice. Thus any errors would be magnified with potentially serious consequences. Though this simple situation could likely be identified and corrected in advanced, more complex situations that can not be anticipated in a dynamic environment would likely occur. The potential occurrence of data looping can best be avoided by including pedigree information (in this case, source and dependency information) in the dissemination of the information.

The challenge with these and similar issues is that the solution is not purely technical. Satisfactorily addressing these issues will require integration of policy, doctrine and technical solutions. Without this effort the DOD will continue to be plagued with inefficiencies and errors, regardless of the architecture implemented in the IC2S.

5. **References**

[McCarthy, *et al*., 1998] USAF Scientific Advisory Board. *Report on Information Management to Support the Warrior*. SAB-TR-98-02, 19 October 1998.

[Kandel, 1992] Abraham Kandel and Gideon Langholz. *Hybrid Architectures for Intelligent Systems.* CRC Press, Boca Raton, FL, 1992.

[Doyle, 1996] Jon Doyle and Thomas Dean. *Strategic Directions in Artificial Intelligence.* ACM Computing Surveys, 1996.