

A Critical Look at Centralized and Distributed Strategies for Large-Scale Justice Information Sharing Applications

A White Paper Prepared by the Integrated Justice Information Systems Institute

Principal Author: Dr. Alan Harbitter, Chief Technology Officer, PEC Solutions

The opportunity for justice information sharing is often thought of in terms of two dimensions: *vertical sharing*, among local, state, tribal, and Federal government entities; and *horizontal sharing*, among first responders, investigators, intelligence analysts, prosecutors, and court and corrections officials. In recent years, the justice community has accelerated activities to improve information sharing in both dimensions.

Improved information sharing throughout the justice community is a national priority. An increasing number of organizations are launching programs to share information and conducting studies to plan for and implement justice information sharing initiatives. This paper adds to the growing body of knowledge by examining system strategies for information sharing. In particular, it looks at the advantages and disadvantages of using centralized and distributed strategies for large-scale, nationwide information sharing.

Successful centralized and distributed justice information-sharing systems are in operation today, and it would be inappropriate to try to select one of them as having the best strategy for all circumstances. Instead, this study attempts to illuminate the positive and negative characteristics of the two alternatives and describe the situations that favor one approach over the other.

A Context for Comparative Analysis

To anchor the analysis and illustrate comparisons, this paper uses an example of a grand-scale justice information-sharing application: interconnecting all police department records management systems (RMSs) in the country. The objectives of such a project might be to allow an authorized law enforcement investigator in Biloxi, Mississippi, for

instance, to gather data about a suspect associated with related incidents in Duluth, Minnesota, and Harrisburg, Pennsylvania, as well as any other locations and jurisdictions that may present themselves during the research process. This paper does not address the feasibility or desirability of actually building such a nationwide network of records systems; the example is used for comparative purposes only.

Centralized and distributed strategies are compared in relation to five qualities of information-sharing systems:

- (1) Cost – It is impractical to synthesize complete system costs for purposes of this example; however, some cost factors cause different results in centralized and distributed systems.
- (2) Governance and data ownership – Governance and ownership policies and agreements are important factors in building information integrity and fairness among those participating in the information sharing community. The alternative strategies present different options for controlling, managing, and establishing policy on information storage and use.
- (3) Performance and function – The purpose of sharing is to provide actionable information to authorized individuals who can make the most use of it. The choice of system strategy directly impacts the type and efficiency of information-sharing functions that can be performed.
- (4) Scalability – Nationwide information sharing involves extremely large numbers of individuals who own or need to access information. While large systems can be either centralized or distributed, the manner and degree to which each strategy can scale up or down is distinctly different.
- (5) Security and privacy – Justice information is typically sensitive. Mechanisms to provide confidentiality, integrity, availability, and privacy vary between the two system strategies.

Distributed System Strategies for Information Sharing

Distributed system strategies spread information across geographic and organizational boundaries. Perhaps the ultimate distributed system is the Internet. The World Wide Web (an application that runs over the Internet) stores and provides access to hundreds of billions of documents containing thousands of terabytes of information. Successful designs for large-scale distributed applications commonly draw from the structure of Internet. The latest buzz phrase in distributed system design is “service-oriented architecture,” which builds on Internet technology, adding new protocols and features that allow the resultant system to do more than simply provide access to multimedia Web pages and support e-commerce.

Figure 1 illustrates a service-oriented architecture in the context of our example – a nationwide network of records management systems. In this example, each records management system owner offers electronic services¹ to the nationwide justice community. Those services may include responding to such queries as, “If I give you a name and a weapon, will you tell me about incidents you know of involving that name and weapon?”

The services that are available are posted in a registry. If an authorized investigator is researching an incident – or, more accurately, if a computer program is researching an incident on behalf of an authorized investigator – that program can see what services are offered and request information using appropriate services. This strategy has several interesting properties:

- It requires well-defined standards for specifying the services, formatting the provided information, and identifying service users to a service provider.

¹ The concept of an “electronic service” is a little esoteric and deserves some explanation. If a computer system offers, for example, to respond to a query for information or to produce an analysis report for another computer system, that could be viewed as an electronic service.

Figure 1. An Example of a Distributed Service-Oriented Architecture

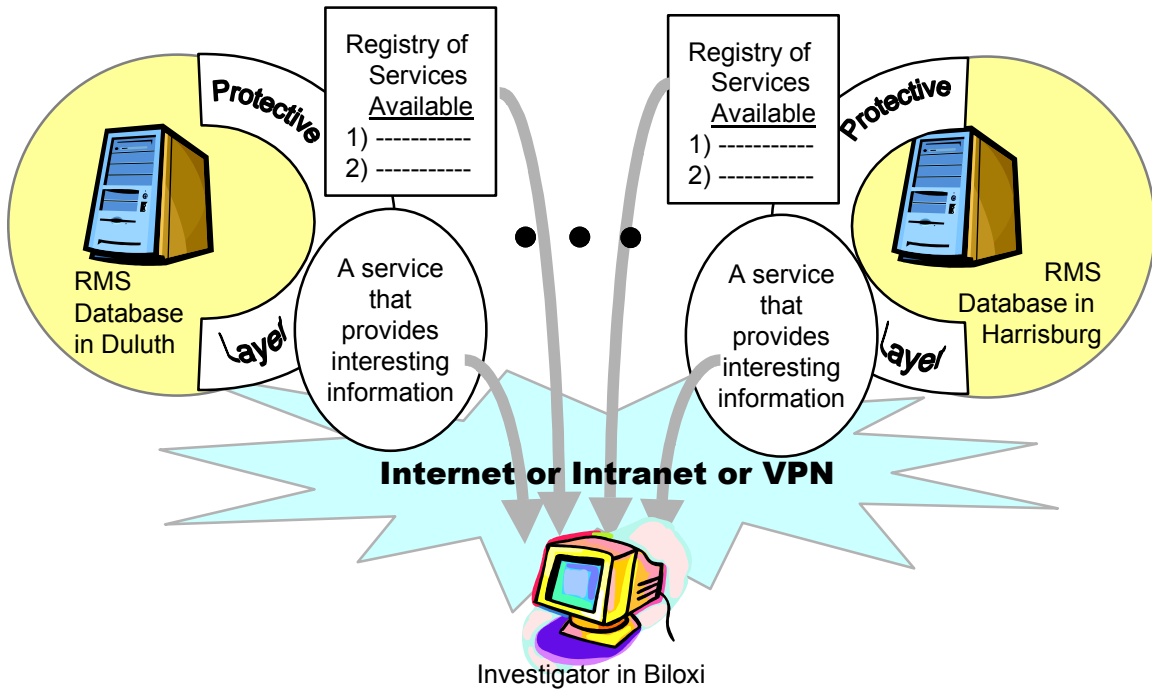
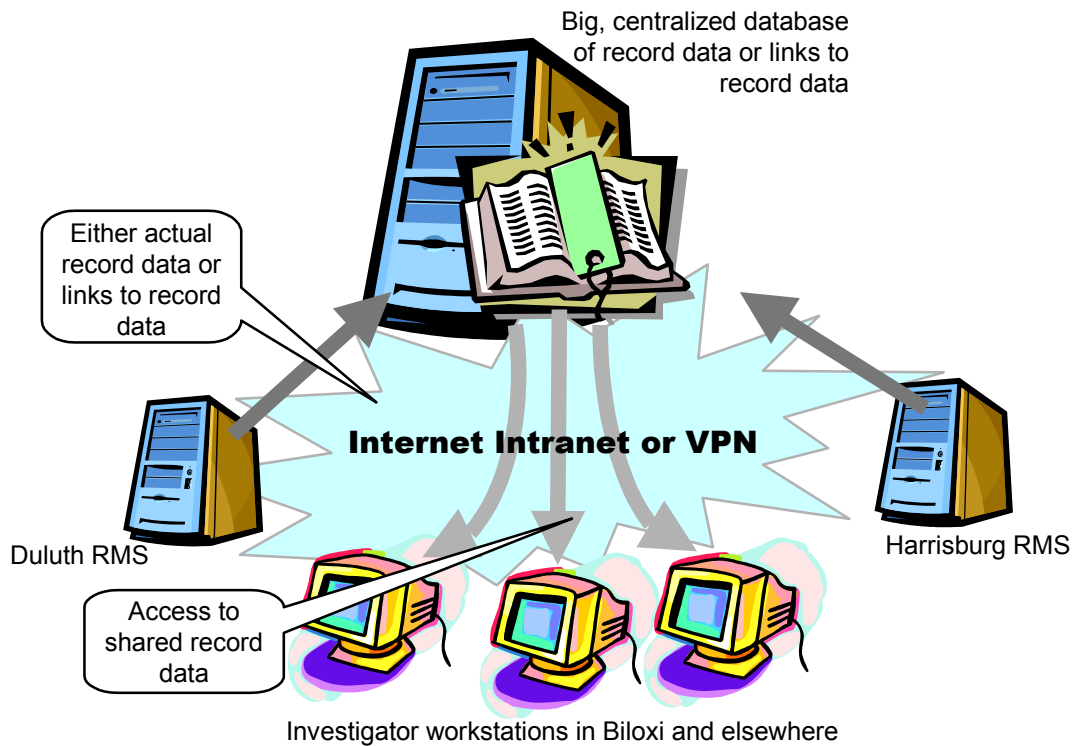


Figure 2. An Example of a Centralized Information Sharing System



- It relies on the willingness of individual system owners to participate in information sharing. Information owners who provide a robust set of services to the community do so by implementing and operating computer systems that can support those services.
- In general, owners of the data can make decisions about what services to offer and what data to make accessible. They maintain control over their information.

Service-oriented architectures reflect the philosophy of the Internet, in which there is a reduced need for day-to-day centralized administration. It has been said that if someone (or some government agency) explicitly launched a project to build the current Internet, that project would fail. The success, scope, and incredible growth of the Internet are a direct result of good technology standards, distributed governance, independence of participants, and strong mutual benefits for participation. These same success factors apply to a large-scale system for justice information sharing.

Centralized System Strategies for Information Sharing

A centralized system places all information in one location. The justice community has many examples of centralized systems. One of the best known is the FBI's Integrated Automated Fingerprint Identification System (IAFIS). The FBI collects fingerprint information from every state in the country and stores it in a huge, centralized database in Clarksburg, West Virginia. A nationwide fingerprint check requires connection to Clarksburg. Similarly, in order to construct a centralized system for the sample nationwide records management project in this paper, a single organization would have to take responsibility for collecting, organizing, and making data available on a grand scale.

Figure 2 illustrates a centralized system using the example of nationwide RMS information sharing. The figure suggests two approaches to centralization. In the first, RMS records are collected, indexed, and stored in a single database, which would have to be big – probably too big to be practical. In the second approach, the size of the database

is reduced by storing central “links” to the information². In this approach, information resides at its source in RMSs across the country. At some point in time, even if only temporarily, the information stored in the local RMSs will be pulled out either by the central system or by the users who want to access that information. In fact, the central system may periodically pull information from multiple local RMSs and perform analyses on the combined data.

The centralized strategy has some interesting properties:

- The currency of the information (or links to the information) depends on the frequency of the collection process. Collection can be a large and logistically complex problem. Even if only the links are stored centrally, they must be updated frequently to reflect the arrival of new data or changes in old data.
- As with the distributed approach, there has to be agreement (i.e., standards) on the format of the data so the information can be meaningfully searched and manipulated.
- Many aspects of the system, such as the services provided, user management policy and procedure, and security mechanisms, can be chosen and administered, almost unilaterally, by the organization that runs the centralized system.

Comparative Analysis

The following paragraphs compare the cost, governance and data ownership, performance and function, scalability, and security and privacy characteristics of large-scale distributed systems and large-scale centralized systems.

Cost. Without making any judgment as to which strategy would result in the overall lower total cost, the following observations can be made about the relative cost characteristics of each:

² This is actually the model adopted in the construction of the Interstate Identification Index (III) operated by the FBI, wherein a central database of links to state criminal history systems and the central system provides a “pointer” to state-maintained criminal history information.

- The cost for the distributed strategy can be shared among the participants. Each information provider can be required to build and maintain systems that provide localized information access and to bear the associated cost of that system. The provision of services may be an add-on to existing local information system capability.
- The distributed system is more amenable to incremental funding because the initial sharing network can start with a few participants and scale up to include the entire community as funds become available. Conversely, in a centralized approach, a large initial investment is required to establish the central repository and develop and implement the structure for data acquisition.

Governance and Data Ownership. While governance and data ownership are equally important and complex in both strategies, the issues are significantly different. Under the distributed system strategy, the owners of the data maintain control and fulfill service requests on a case-by-case basis. Under centralized strategies, the responsibility for – and, to some degree, ownership of – the data is transferred from the original owners to centralized service providers when the data (or the links to the data) are collected.

The participants in a centralized system give up some degree of control over the distribution of the contributed information. This has traditionally been a cause for concern on the part of criminal justice executives. From a political perspective, this perceived loss of control has discouraged the creation of centralized systems, particularly across multiple disciplines in the justice system. Justice executives are often more comfortable with a distributed system concept in which they believe they can control dissemination and retain the ability to disengage in information sharing.

The focus of governance in a centralized sharing strategy is to set policy and to manage:

- the collection of information from sources and the guidelines for access and use of shared information, and
- the protection of the security and integrity of the collected information.

The focus of governance in a distributed sharing strategy is to set policy and to manage:

- the standards each information provider uses to impose consistency on the information provided and the manner in which it is provided,
- the rules for participation that convince information owners their information is not being abused and convince information users the information is dependable and accurate, and
- the aspects of the system that continually assure participants that there are benefits to remaining an active member of the information-sharing community.

Under either strategy, governance must include all participants in the information-sharing community. Successful governance under the distributed strategy requires coordination among many different peers, each of whom must believe there is personal benefit in complying with agreed-upon policy. Successful governance under a centralized approach is more hierarchical in nature and requires that the organization running the central system have authority to create rules and policy with which information source providers will comply.

Performance and Function. The easiest way to illustrate the differences in performance and functionality between the two strategies is to think about how to implement two common functions performed in information sharing: data collection and query.

While information collection is a critical function in the centralized strategy, it is nearly absent in the distributed strategy. The efficiency of data collection determines the accuracy, currency, and completeness of information. In general, it will not be possible to collect as much information and to keep it as current in the centralized strategy as compared with the distributed strategy, in which users are motivated by their own discipline, policies, and requirements to keep information current.

Whether it is the data or links that are centrally stored, the centralized system supports query and search more effectively than the distributed system. Querying and searching of the distributed system require knowledge of information location and may require an electronic “visitation” to many different sites to satisfy a single information request.

For an interesting view of search issues in centralized and distributed systems, consider the Internet-based music and media sharing networks, “Napster” and its more recent

counterpart, “KaZaA.” Participants in the Napster community – specifically the circa 2000 Napster that was shut down by a Federal court – stored music files on their PCs, but allowed index information to those files to be stored *centrally* on Napster servers. A subscriber interested in locating and downloading a particular song would search the Napster index to locate instances of the song stored in the distributed network. The target file would be transferred directly from the PC on which it was stored to the PC of the user looking for the song. A Federal court ruled that Napster was facilitating illegal activity by storing indices to copyright-protected material.

Conversely, current media-swapping network systems such as KaZaA do not maintain a central directory of file locations. The directory is distributed to member PCs along with the media files. Spider-like song searching algorithms electronically transit from PC to PC on the Internet, looking for a specific title or for PCs that “know” where the title might be located. This dodges the court ruling, but it is an ineffective way to search for specific songs.

While the Napster/KaZaA example is significantly different from this paper’s nationwide RMS example, it aptly illustrates some of the functional limitations of highly distributed information sharing.

Scalability. Although some very large centralized systems are scalable, the distributed strategy has a clear advantage in the area of scalability. No centralized system rivals the size of the ultimate example of distributed information sharing: the Internet. The growth pattern and rate of the Internet is also a testimony to the scalability of distributed information systems.

Security and Privacy. Under the centralized strategy, a large amount of valuable information is stored in one place. Initially, it may seem that the centralized system poses a security liability by presenting a clear target, which once compromised, exposes everything to a would-be cyber intruder. However, with the proper investment in security technology and procedures, this central cache of information can be adequately protected against compromises to confidentiality, integrity, availability, and privacy.

In the distributed strategy, it is a more complex matter to ensure that each information-sharing participant provides sufficient protection. In general, the technology for

protecting information and identifying authorized participants in a highly distributed network is not as mature as the technology used for centralized information organization and sharing. While Internet standards groups such as OASIS and W3C are rapidly developing new protocols, technologies, and standards, it is a technically complex problem to identify and authorize users and protect information in a distributed large-scale information-sharing environment.

Achieving security and privacy objectives in highly distributed systems depends on establishing electronic trust among the many participants – both those who provide and those who request information. This includes, for example, implementing mechanisms that guarantee that a person or software program requesting information has been authenticated and authorized by a process that has been accepted by the agency owning the information. There is no national model for building this type of electronic trust, and such a model is sorely needed.

Establishing rules for and protecting the privacy of individuals whose names may be entered in individual RMSs also varies with the chosen strategy. It is easier to enforce rules relating to dissemination, purging, sealing, or correcting information if all aspects of data management are centralized; it is harder to ensure that all entities participating in a distributed system adhere to common principles of privacy protection. In contrast, the perception by legislative bodies and the general public regarding the extent to which individual privacy rights are supported differs. The general reaction to centralized databases is that the mere aggregation of information has a greater potential for violating privacy rights than a distributed system. Court case rulings have supported this view. Instances in which states have withdrawn from systems or shut down their own intelligence systems resulted primarily from this common perception. The distributed system with local ownership of data is generally perceived as a lesser threat to privacy.

Conclusions

When asked, “What is a better strategy for large-scale information sharing – centralized or distributed?” the definitive conclusion is, “It depends.” The purpose of this white paper is to provide a brief discussion on the factors influencing such a decision. If it were necessary to be summarize with a few rules of thumb, the following might serve:

- If the problem is to share information among many, many systems, the Internet and highly distributed systems strategies such as those envisioned by service-oriented architectures provide a proven, successful model.
- If there is a requirement for complex data analysis and quick, system-wide query, the centralized strategy has distinct advantages. The security issues are also easier to manage in a centralized strategy.
- Distributed system models allow information owners to control their information. This issue has historically been of significant concern in the justice community for both governance and privacy reasons. The controversy over information location and control in centralized systems can be tempered, to some degree, if links to information (rather than the actual information itself) are stored in the central database.
- A distributed strategy allows information-sharing system implementation to begin small and scale up.

While there is a role for both centralized and distributed information sharing in justice applications, information control, privacy, and funding issues favor distributed approaches in large-scale applications.