12th ICCRTS

"Adapting C2 to the 21st Century"

Title: "Do You Know Where Your Information Is? Information Asset Exploitation Across Large-Scale Distributed Enterprises"

Topics: C2 Technologies and Systems, C2 Concepts, Theory, and Policy, Network-Centric Experimentation and Applications

Authors: Sekar Chandersekaran, Institute for Defense Analyses

Andrew Trice, Institute for Defense Analyses

Point of Contact: Andrew Trice

Name of Organization: Institute for Defense Analyses

Address: 4850 Mark Center Dr., Alexandria, VA 22311

Telephone: 703-933-6543

E-mail Address: atrice@ida.org

Do You Know Where Your Information Is?
Information Asset Exploitation Across Large-Scale Distributed Enterprises[1]

Abstract

In the emerging net-centric environment, a key concern is how to achieve efficient information discovery in the face of increasing proliferation of both structured and unstructured information assets across the DoD enterprise. Another critical consideration is managing community of interest (COI) content and vocabulary within a federated architecture. This paper outlines the emergence of a new environment for information asset exploitation within the USAF that addresses some of these issues and enables more effective and information enterprise discovery. To implement this effectively, a range of application capabilities and supporting infrastructure services must be provided.

The paper begins with a motivating discovery example and a discussion of the enterprise view of the overall environment in which the discovery and federated search take place. We next describe a methodology for scenario analysis used to derive more refined enterprise requirements, and detail the emerging solution for information asset exploitation. Finally, we summarize the benefits of the solution and provide status on its development.

[1] The publication of this paper does not indicate endorsement by the Department of Defense or IDA, nor should the contents be construed as reflecting the official positions of those organizations.

I. INTRODUCTION

As the amount of information generated by multiple military organizations and programs continues to proliferate, efficient information discovery becomes more and more problematic unless sufficiently powerful enterprise infrastructure and solutions for discovery can keep up. For instance, in the USAF, there are many systems and data repositories spread across many different major commands and programs. While there have been some successes in integrating and deploying certain USAF enterprise systems, achieving a much greater degree of information access over a wider variety of information assets across the USAF (and other services as well as agencies) remains a key goal in the path to net-centricity in command and control as well as supporting functions.

Another critical consideration is managing community of interest (COI) content and vocabulary within a federated architecture. On the one hand, COIs are responsible for identifying authoritative content and the vocabularies and other metadata that describe that content within a prescribed domain, which should increase the consistency and quality of discovery results. On the other hand, users often wish to perform discovery across multiple COIs, which immediately raises many questions about the effectiveness of federated search across COI vocabularies and the adequacy of mechanisms for cross-domain information asset access and security.

This paper outlines the emergence of a new environment for information asset exploitation that addresses some of these issues and enables more effective and information enterprise discovery. First, we provide an example of a typical information discovery scenario and its potential complexities in a large enterprise. Second, we discuss the enterprise view of the overall environment within which discovery must take place. Third, we describe a methodology for scenario analysis used to derive more refined enterprise requirements. Fourth, we detail the emerging solution for information asset exploitation within this environment, considering both the vocabulary and enterprise issues. Finally, we summarize the benefits of this approach and provide status on its development.

II. EXAMPLE: DEPLOYMENT READINESS

One of the objectives of the USAF is to provide better information on individual readiness to deploy for a specific mission. This can include many functions, such as managing readiness requirements, tracking individual (medical, equipment, training, administrative and legal) readiness item status, evaluating individual readiness status, facilitating readiness (scheduling) actions, and monitoring and reporting individual readiness status consistently across the Air Force.

Over a particular mission thread such as Air Operations, planners need to quickly understand readiness of personnel and staffing options. Moreover, most users will need to be able to perform this task without having extensive knowledge about where the information resides, format(s) in which it is represented, or how it might have to be

aggregated. This is in contrast to certain categories of users who might be willing and able to understand and manipulate the structure of the data (e.g., experienced intelligence analysts).

In this context, a planner may wish to pose a "straightforward" query such as the following:

Find personnel summaries of all individuals with

- Skills A, B, C
- Experience in missions of type D, E
- Medical fitness level M
- Security Profile X
- Availability within the next 10 days

While this query may (and should) seem to be conceptually straightforward to the planner who is the end user, varying degrees of enterprise complexity may need to be addressed to execute such a query. Most generically, the query could be a sophisticated federated search.

First, the information relevant to the above topics for a particular activity (say, mission planning) could be spread over disparate domains. For example, the personnel skill, experience, and availability information, along with the personnel summaries themselves, may exist in one COI, the medical information may exist in another, and the security information in a third. Moreover, these COIs may physically exist in different enterprise domains or "enclaves" (see below). This introduces some complexity in federated search, but does not even begin to consider any potential access restrictions that must be enforced, either by the COI, by individuals, or by other business rules (e.g., privacy or HIPAA rules). A comprehensive solution to this set of issues must address all these complexities effectively.

III. ENTERPRISE VIEW AND OVERALL ENVIRONMENT

Before we begin describing the details of Information Asset Exploitation it would be helpful to describe the nature of an enterprise from a computational perspective. It will be followed by a generic model for Information Asset Exploitation that will be the goal of the architecture.

Most enterprises usually consist of a number of enclaves (Figure 1). Each enclave is a collection of users, service requestors, service providers, networks, routers, firewalls, server clusters and many different types of devices. There is a logical encapsulation of all these entities within the enclave and the enclave manager(s) are responsible for managing all these entities. For example, the identity manager is responsible for issuing credentials to all the entities, the configuration manager will manage the versions of all components running on all entities; the system health manager will monitor the health of all components (hardware and software) within the enclave. Also all the entities within an enclave use the same type of credentials, adhere to the same set of security policies, trust

one another's credentials and also adhere to the same management policies. The platforms may be different: some may be Windows, others may be Unix, etc., but the assumptions made above will still be valid.

Where there are multiple enclaves, the enclaves are related to one another by means of specific "trust relationships" agreed to between the management of the different enclaves. As an example, not all members of an enclave will be allowed access to services in a foreign enclave and in an extreme situation they or programs running on their behalf may not even be able to see the existence of some enclaves. The rationales for the existence of multiple enclaves are many. For example it may reflect a business model where an enterprise might consist of autonomous businesses with separate policies and profit / loss objectives for each enclave. Another reason might be the nature of the enterprise came about due to mergers and acquisition of multiple businesses etc. In a similar fashion the same model can be expanded to incorporate partner enterprises or collaborating enterprises. The nature of collaboration can be tightly controlled by the trust policies set up between the enterprise managers. Yet another reason might be that some enclaves run at one security level [in DOD parlance] while others run at a different security level. These may be NIPRNET-based enclaves or SIPRNET-based enclaves.

It is important to realize that such an enterprise model fully supports net-centric computing as well as a Service Oriented Architecture whose scope runs across multiple enterprises [1,2].

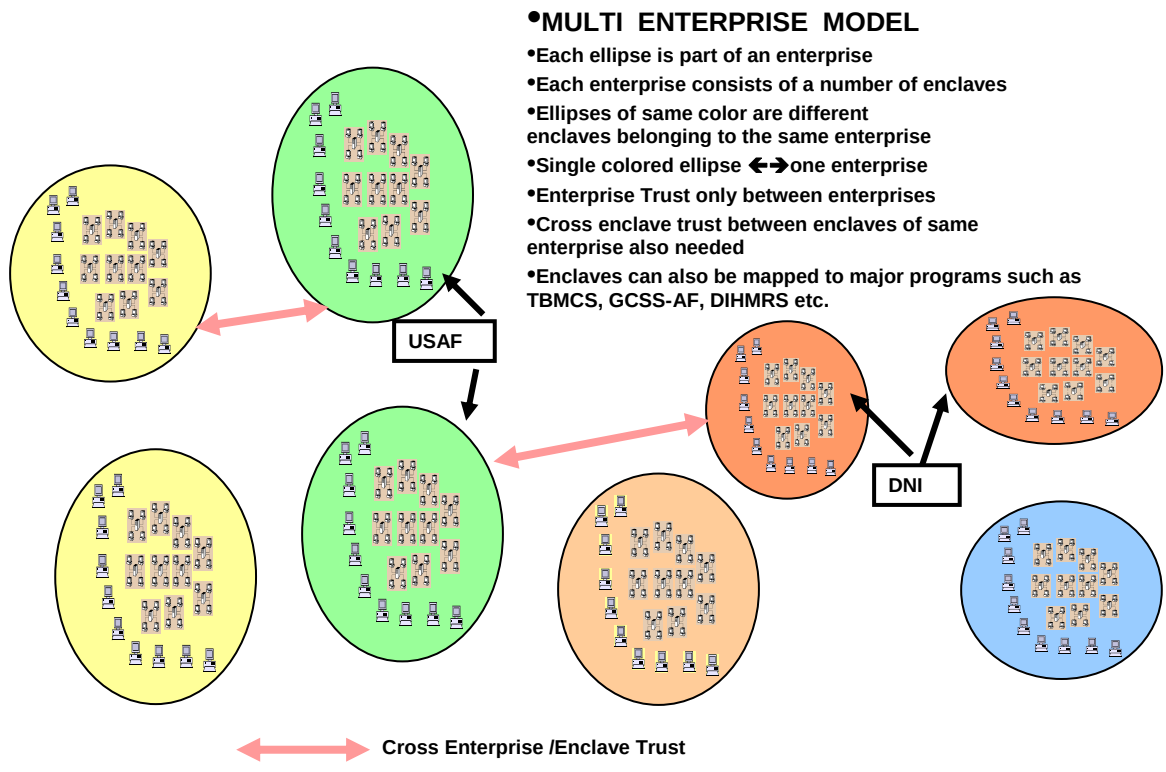

Figure 1: Multi-Enterprise Model

Once the enclaves are defined and set up, one might consider the shape of the specifics of the Information Asset Exploitation model (Figure 2). The goals here are the realization of net-centric computing while using a service oriented architecture. So the canonical model assumes that requestors are in any enclave and that services providing access to metadata registries as well as information content may be in any enclave.

The request may initially traverse to a local service which (possibly) redirects it to a service in the same enclave or to a service in a different enclave. So a query requesting readiness of components of various categories may find itself traversing many different enclaves seeking information asset components that are located elsewhere. These components are possibly aggregated by the initial service before returning it to the original requestor.

For the cascading of such requests to work in a heterogeneous environment, federation at many layers will have to be worked out (Enterprise Service Busses, Name federation, Credential federation, Discovery federation, Workflow federation [3]). These aspects are not discussed in this paper.

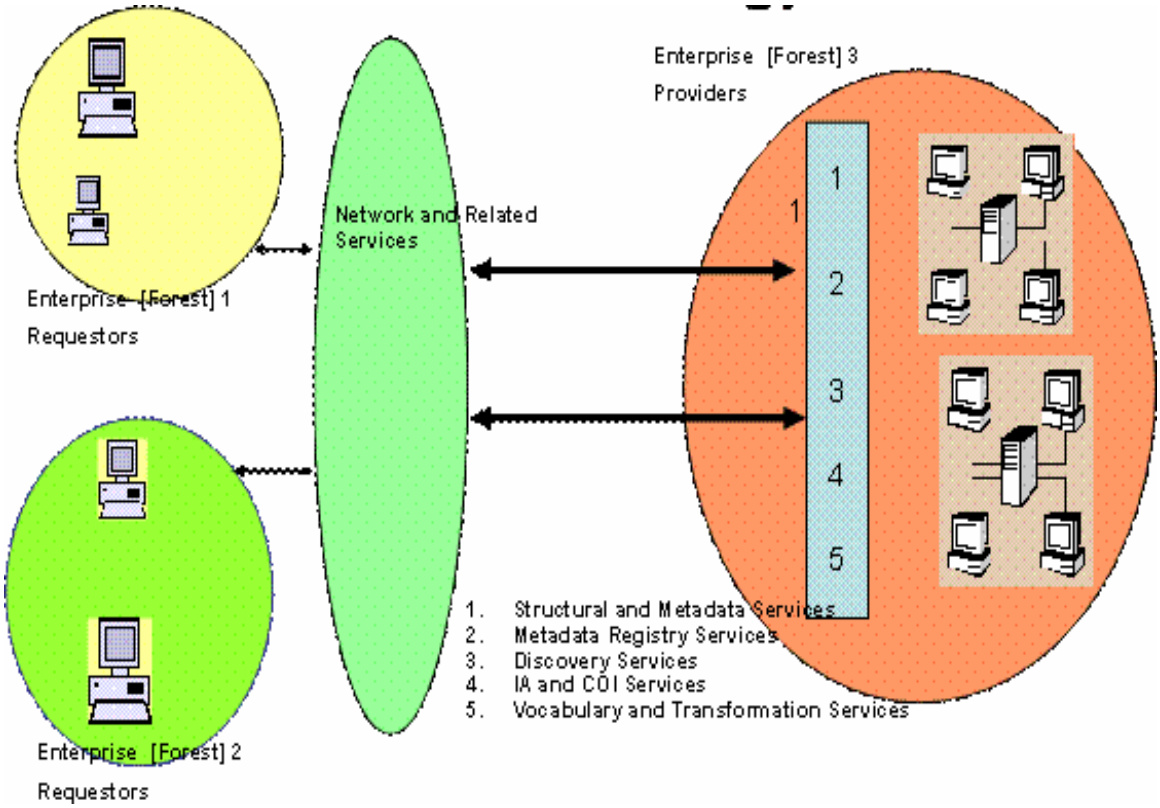

Figure 2: Model for Information Asset Exploitation

IV. SCENARIO ANALYSIS

To bridge the layers between the functional requirements of an information asset exploitation environment and the supporting enterprise and infrastructure requirements, we performed a detailed analysis of the scenarios provided from a conceptual requirements document. These scenarios described what must happen from a user point of view, but were not intended to address the overarching enterprise, infrastructure, and security context within the user requirements must operate. Defining the larger context and providing greater detail and analysis is very important for understanding enterprise requirements and writing system specifications at a lower level.

The scenario analysis involved creating matrices that take each user-level step and translate it into one or more (possibly more detailed) steps that incorporate the larger context noted above. For each step, we identify several critical dimensions that help to refine the functions needed, identify open issues, and begin to identify services required in support of the function. Table 1 shows an excerpt from the matrix for a "content consumption", or end user discovery, scenario analysis.

No.	Step	Output	Preconditions	User-Level Questions	System-Level Questions	Supporting Services (Auth., Athzn needed for all)
1	User logs on / authenticates	User's "landing page" or starting page in application	-User has appropriate credentials	What range of authentication methods will be supported?	How to handle Windows forest account vs. non-Windows accounts?	Credential, Audit, Timestamping, Device Services
2	User connects to COI list and discovery "portal"	List of relevant COI's	User is member of at least 1 COI; Assumptions (Stage 1) COIs in 1 enclave (Stage 2) COIs in multi-enclaves	How does user know and select COI's?	How has system determined and validated user COI membership? Who creates COIs and members?	COI Retrieval and Validation, Data
3	User requests list of available tailorable queries	List of tailorable queries	User has validated system about COI membership, If User environment not a web page, app needs to be launched	How does user know about / access web page or application containing queries? Are the query pages common across COI's?	What is the basis for selecting query pages? Is the query related to the metadata structure? Who has constructed and stored the queries?	Metadata, Data, Discovery, Naming Services

Table 1: Scenario Analysis Excerpt

The original user scenario was concerned with the discovery of content, which could be structured (e.g., result from a database query) or unstructured (e.g, a text document). This particular scenario analysis considers not only the steps specific to the discovery (e.g., performing searches), but also the steps preceding the search itself such as login and entry into an application from which the discovery takes place for the end user.

For each step in the scenario, we identify the following components.

1) The step number and name.

2) The expected output from the step. This defines what information or results are available to the end user or to the system for subsequent steps.

3) The preconditions for the step. This represents what data or conditions we must be able to assume exist prior to the step being completed. These preconditions can often identify the need for other scenarios and capabilities.

4) User-level questions. This identifies any unresolved issues around the user experience, access to information, or the larger business process context of the step.

5) System-level questions. This identifies any unresolved issues around infrastructure, security, deployment, or other technical considerations.

6) Supporting services. This defines the enterprise services needed to support the activities in the step. Note that Authentication (Auth.) and Authorization (Authzn.) services are assumed to be necessary throughout the scenario.

The scenario analysis is a valuable tool for identifying additional functionality not identified in conceptual requirements, and ultimately for identifying functional, enterprise and infrastructure services.

V. EMERGING SOLUTION: INFORMATION ASSET EXPLOITATION ENVIRONMENT

The USAF's proposed solution to problems like the above is an Information Asset Exploitation Environment (IAEE). The IAEE is premised on the following key features:

1) Authoritative vocabularies identified by empowered COIs and by senior USAF leadership, with lines of responsibility between COIs clearly understood and managed, and mapping/alignment of vocabulary across COIs clearly defined. This results in improved precision (proportion of returned results that are truly relevant) and recall (proportion of total relevant results that are returned) for information discovery.

2) Authoritative sources for information assets (both structured and unstructured) provided by COIs that can be collected, indexed, and tagged with the corresponding vocabulary and subsequently discovered and exploited by authorized users. COTS indexing and search tools are used to tag and retrieve the information assets. Systems not providing authoritative sources can be potentially decommissioned if they contain only duplicate data.

3) A runtime environment with the following four major components: a) Collections of information assets; b) a metadata registry describing the vocabularies and other key types of metadata, consistent with DoD and USAF standards; c) a metadata catalog linking instances of information assets to its associated metadata tags and associated access service; and d) a service registry describing how information assets can be accessed.

4) A federated enterprise environment that enables discovery and secure information sharing across USAF domains or enclaves as outlined above, using cross-domain trust facilities offered by standard operating systems interoperating with COTS tools (e.g., indexing and search tools, service registries) via industry standards (e.g., OASIS, W3C) in an SOA environment.

In the example deployment readiness query introduced in Section II, the above features of the IAEE can be applied to its various potential complexities as follows.

If the query must be implemented as a federated search across multiple COIs, the IAEE must reach across multiple COI metadata registries and multiple metadata catalogs to retrieve the data. For a user residing in a given enclave, cross-enclave trust is used to enable them to access the metadata registry or information access in another enclave.

VI. SUMMARY

The key benefits of the IAEE include improved precision and recall for information discovery, long-term cost savings due to elimination of redundant systems and repositories and leveraging of COTS tool and standard operating systems rather than custom solutions, and leveraging of information assets across the federated USAF enterprise. The IAEE is currently under pilot development for a key COI domain operating in a single enclave; as authoritative vocabularies in other COIs continue to be developed and additional components are deployed the IAEE will be expanded to other enclaves and ultimately the entire USAF enterprise.

VII. REFERENCES

[1] AFCA/ESCO Lead Command, “Air Force Directory Services Overview”, USAF briefing at the Service Oriented Architecture Working Group, 04 Jan 2006.

[2] Robert Bogue, “How SharePoint and SOA Fit Together”, Intranet Journal, 3/16/2006. (http://www.intranetjournal.com/articles/200603/ij_03_16_06a.html)

[3] Microsoft Corporation, “Advanced SharePoint Workflow Using Custom Document and Form Libraries. (http://whitepapers.zdnet.com/webcast.aspx?&docid=158607&promo=100511)