Foreign Media Collaboration Framework: A Service-oriented Architecture for Linguistic Translation and Processing

Garin R. Clint and Richard J. Thomas
Northrop Grumman Information Systems

Automated linguistic translation and processing are playing an increasingly critical role in both the public and the private sectors, filling in where human translation or transcription is impractical, inefficient, or cost-prohibitive. Users need a service-oriented solution that integrates a family of complementary products and technologies to provide machine translation of speech, text, and media broadcasts in a scalable, highly adaptive architecture supporting and supplementing available linguists, as well as enabling nonlinguists to work and collaborate at productive levels.

To address that need, Northrop Grumman Information Systems is developing the Foreign Media Collaboration Framework (FMCF), a service-oriented architecture that integrates and manages the application of the foreign-language translation capabilities of the best machine translation, optical character recognition, and automatic speech recognition software available today. As a service-oriented capability, the FMCF can satisfy user needs by offering intrinsic interoperability, increased federation and vendor diversification, better alignment with business and technology domains, greater return on investment, and improved organizational agility. This article provides an in-depth exposition of FMCF concepts and methodologies, as well as a much-needed knowledge base for foreign media collaboration problem solving.

Introduction

Since the 1990s, intelligence community (IC) organizations such as the National Air and Space Intelligence Center (NASIC) have embraced machine translation (MT) and other linguistic processing technologies to fill the gap between the limited number of expensive and overburdened human translators and the vast quantities of data requiring translation and preliminary analysis. Recently, however, those technologies have advanced rapidly, fueled by intensifying demand for automated linguistic processing, primarily in MT, optical character recognition (OCR), and automatic speech recognition (ASR). Governments and the IC are no longer the sole source of research and development (R&D) funding in this arena. Major corporations are also investing millions of dollars in their own development projects and internationalization, further driving the state of the art for related technologies, tools, and applications.

In response to the growing interest in MT alone, the National Institute of Standards and Technology's (NIST's) 2006 Machine Translation Evaluation [1] analyzed 40 participating organizations, including current-generation commercial MT vendors; the research arms of information technology leaders such as Microsoft Corporation, Google, Inc., and IBM Corporation; universities; and government organizations from nine countries. All participated in the event with their latest rule-based, statistics-based, and hybrid technologies.

Given the mix of individual strengths and weaknesses of each of the diverse vendor tools, the full suite of all linguistic processing technologies available today offers an impressive array of capabilities. Those technologies are already used throughout the IC, as well as a range of organizations in the public and private sectors. Even individual home users have ready access to Web-based and desktop tools. The continued demand for and infusion of R&D dollars by both government and industry promises highly enhanced capabilities for the future.

No single vendor, however, can currently provide the most advanced technologies for all languages and means of language translation. To meet diverse translation needs, a combination of vendor tools and applications is now and will continue to be required. Therefore, Northrop Grumman Information Systems is developing the Foreign Media Collaboration Framework (FMCF), a service-oriented architecture (SOA) that integrates and manages the application of the foreign-language translation capabilities of the best machine translation, optical character recognition, and automatic speech recognition software available today. Our experience to date with the FMCF demonstrates that leveraging multiple complementary vendor tools and capabilities, delivered as services to the user in an integrated, net-centric fashion, is a best-of-breed approach to providing translation software capability modules that are scalable, tailorable, interoperable, user friendly, easily deployable, and supportable.

Problem: Enterprise Foreign Media Processing

The Department of Defense (DoD) and the IC's foreign media processing requirements include

- Speech-to-speech, text-to-text, speech-to-text, and text-to-speech translation capabilities in numerous priority languages
- The ability to perform diverse missions, many requiring specialized dictionaries

Dozens of individual vendor solutions may be combined to provide complete coverage for all required mission applications and leverage the best MT, OCR, and ASR tools and technologies available. In addition, significant cost and technical risk may be encountered when individual vendor solutions are integrated into a user's automated business processes and workflows.

Such integration is no simple task, because each vendor typically uses proprietary interfaces, data models, domains, and naming conventions. The underlying technology and platform support requirements can also differ considerably, as can vendor capabilities. Some vendors support a number of languages across multiple subject domains with varying degrees of success, whereas others focus on just one language or subject domain. With the insertion of new features and algorithms, the application programming interfaces (APIs) for even a single vendor can change significantly from one software release to the next. Except for the Translation Memory eXchange (TMX) eXtensible Markup Language (XML) format, interoperability standards between vendors currently do not exist. Those factors, combined, result in high integration cost and technical risk with each new release of any component technology or tool.

Buying, supporting, maintaining, and training a diverse set of tools from a diverse set of vendors can be both complex and costly:

- Some high-end MT applications can cost more than $200,000, excluding hardware.

- ASR hardware and software prices may run an order of magnitude higher than those for MT-related items.
- Deployment, upgrade, support, and training of individual vendor tools and applications are also expensive.
- Significant ongoing maintenance charges can arise as information technology staffs administer disparate linguistic applications, and add or replace individual tools and application modules with newer technologies.

The DoD and IC seek to close critical language translation gaps—shortage of translators and limited ability to coordinate with coalition partners, monitor/evaluate civilian information, and communicate with foreign language speakers whom soldiers encounter during missions. Linguists or translation capabilities will be needed wherever warfighters come in contact with or must coordinate with foreign speakers. Large numbers of sound files and documents must be triaged and analyzed to help focus and prioritize the work of available linguists. Capabilities must be net-ready to support net-centric operations, maintenance, upgrades, and embedded training. Northrop Grumman's FMCF provides a tested, scalable SOA to integrate and deliver the wide variety of required vendor tools and applications [2].

Solution: Enterprise Service-oriented Architecture Framework

The FMCF offers foreign-language translation capabilities as net-enabled services in a SOA framework with standardized front-end user interfaces. That framework can significantly reduce the cost and risk of integrated capability delivery, maintenance, and upgrade, as well as streamline both user access and training. The key elements of the FMCF solution are

- The logical, physical, and functional characteristics of the layers
- The interactions between the layers
- The relationships between the layers and the users—both consumers and producers

The FMCF consists of four logical layers, shown in Figure 1, that support both consumers and producers of the framework's resources:

- The *consumers* include linguists, translators, and analysts.
- The *producers* are the software developers and system integrators.

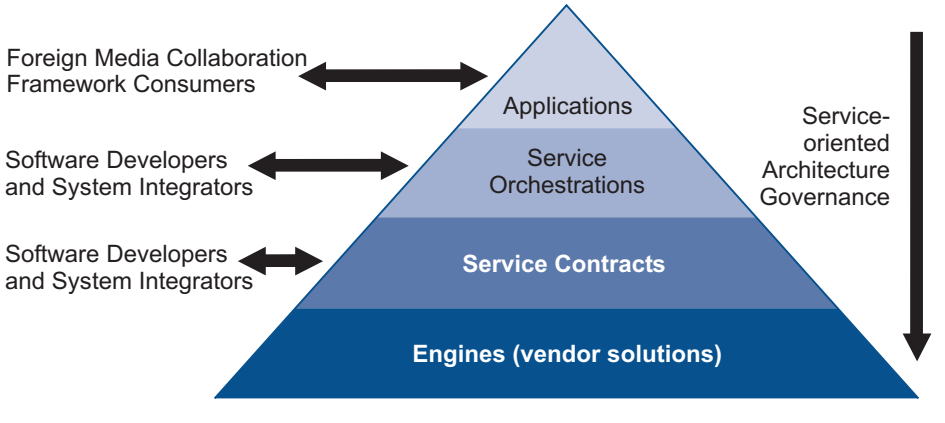


Figure 1. Foreign Media Collaboration Framework logical view

The *applications* are the interfaces through which FMCF users interact. Either existing Northrop Grumman-built applications or third-party applications may be used. Northrop Grumman Information Systems has created several interfaces to the FMCF, including the FMCF Portal, the Translator Workbench, and the Digital Library Input Processing System (DLIPS). All are detailed in the subsection titled "Existing Products" (pages 65–67).

The Applications layer interacts with either the Service Contracts layer or the Service Orchestrations layer. The Service Contracts layer is used by software developers and system integrators to either develop a core service or access an existing service.

A *service orchestration* is a set of service contracts that are connected to each other in a workflow to meet specific business objectives. When a service orchestration is developed, software developers and system integrators can then use the orchestration in their applications to perform complex processing. Further, service orchestrations that comprise other service orchestrations can be built.

Most primitive is the Engines layer, providing the interface between a service contract and a vendor engine—which may be commercial off-the-shelf, government off-the-shelf, or designed specifically for the FMCF.

Figure 2 depicts the FMCF layers in a physical context. The framework integrates capabilities from multiple vendors, allowing FMCF consumers to use limited resources efficiently. The FMCF is designed to communicate with other FMCF nodes located on the same physical wide-area network (WAN), such as the Internet. A node is a physical implementation of one to many FMCF servers on a local area network (LAN) that operate as a single system. The framework abstracts vendor capabilities behind standardized service contracts, thereby driving reductions in total cost of ownership. Further, the FMCF reduces the need for multiple instances of the same vendor capabilities in multiple locations and configurations. When any user interacts with FMCF services, the FMCF identifies the most suitable resource across the entire enterprise (all nodes) to meet that specific user's needs.

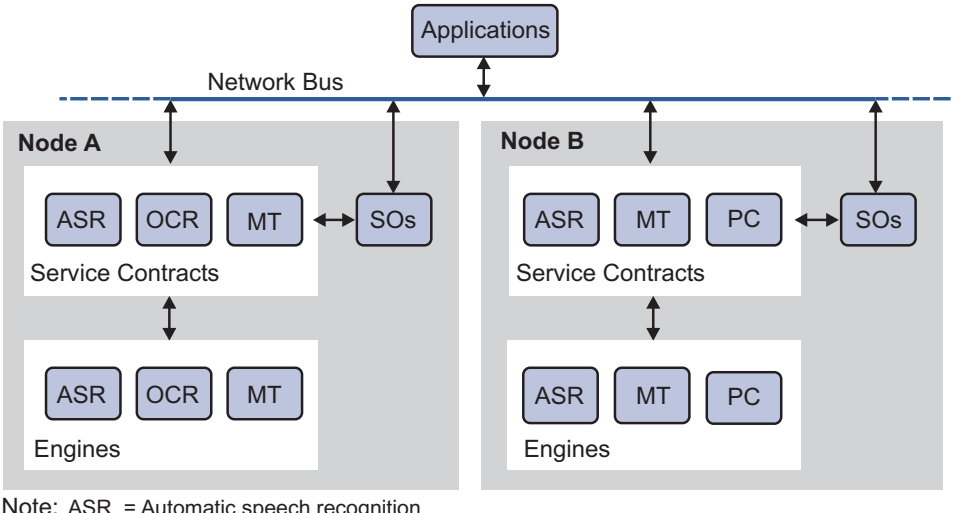
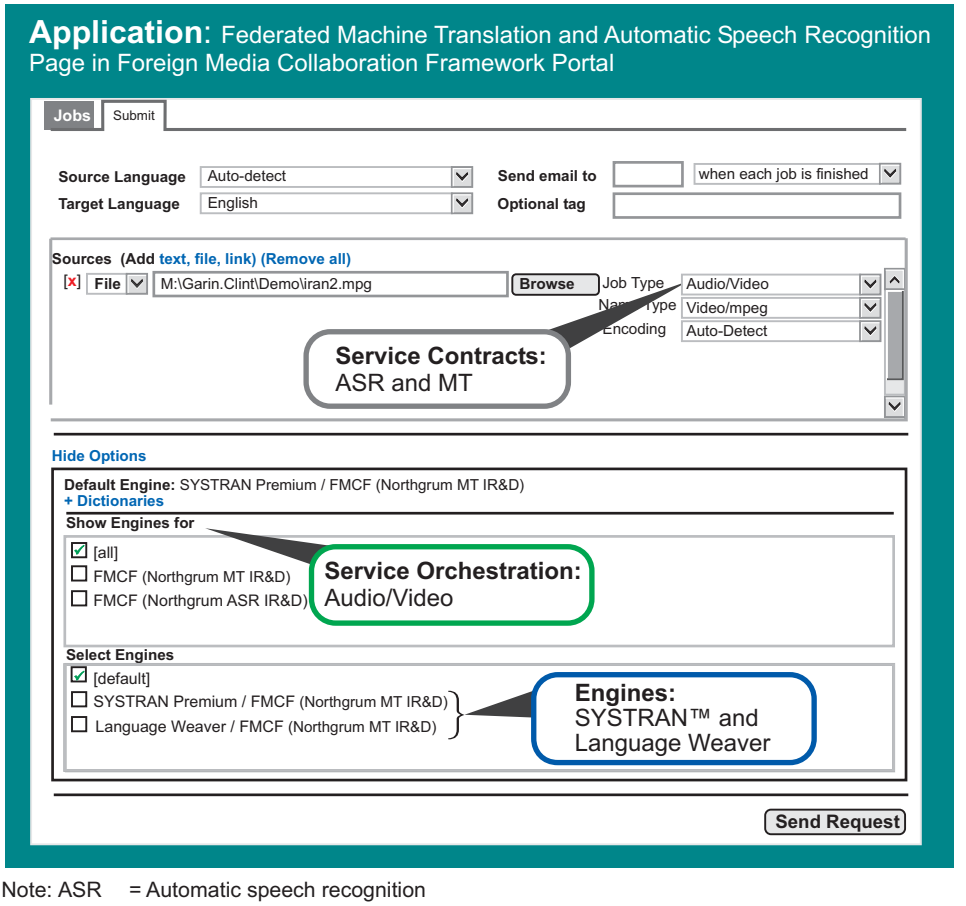


Figure 2. Foreign Media Collaboration Framework physical view

A third way to visualize the FMCF layers is in a functional implementation, illustrated in Figure 3. The example task is the processing of a video file containing foreign content and requiring, as output, a video file in another language. The functional capability includes an application and a service orchestration. The service orchestration, in turn, requires a video upload, ASR, MT, and product generation in a video format. Figure 3 displays the Federated MT and ASR interface, which is built into the FMCF Portal. That application enables FMCF consumers to translate Web pages, office products, and video files.

The complete functional view includes the processes involved in implementing a capability that fully meets a user's needs. The FMCF is flexible, accommodating a wide range of translation capabilities. The key to flexibility is the ability for any user with development support to create a new capability to meet specific needs. Figure 4 shows the steps taken by a new user, beginning with the definition of requirements. Based on those requirements, the user first initiates a search and evaluation process, via the appropriate applications, and then determines whether the desired service already exists somewhere in the enterprise network. If existing capabilities do not fully satisfy the user's requirements, the



Note: ASR = Automatic speech recognition
FMCF = Foreign Media Collaboration Framework
MT = Machine translation
IR&D = Independent Research and Development

Figure 3. Foreign Media Collaboration Framework functional view

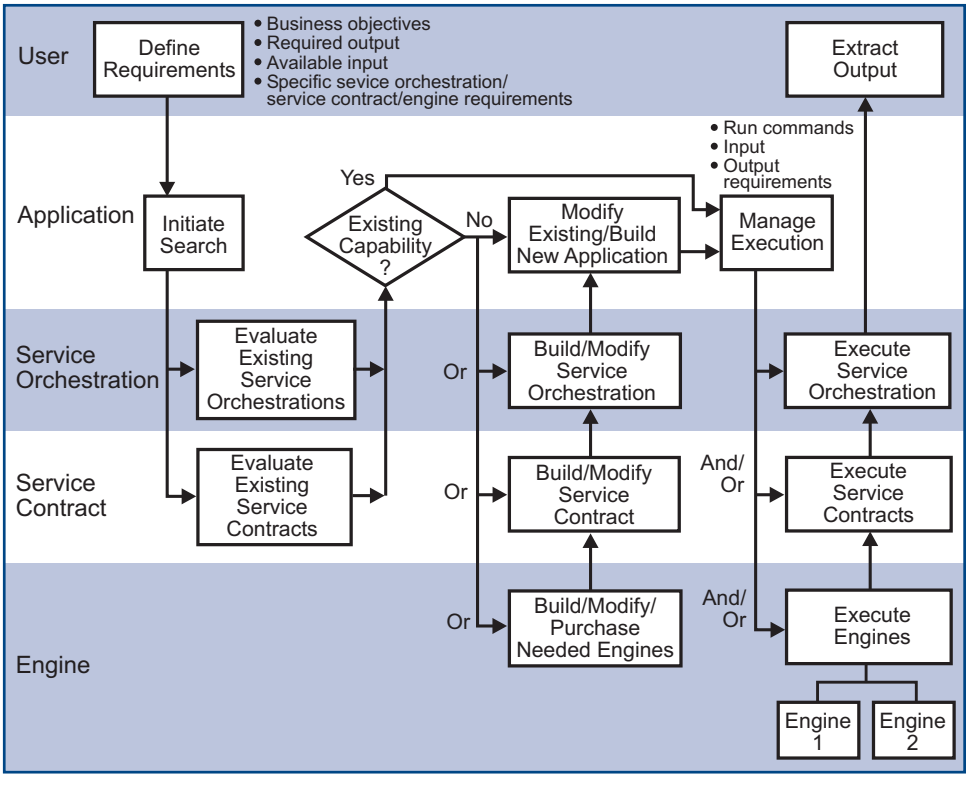


Figure 4. Foreign Media Collaboration Framework functional flow

existing applications, service orchestrations, service contracts, and/or engines must be modified or new ones must be added.

Input is managed by one or more applications and inserted into the processing at any level from service orchestration to direct operation of one or more engines. The user and the development team can build a fully automated capability or allow for user intervention at any level. The output also is managed by the application(s), providing reports and products, as specified by the user. As with the main processing functions, user interaction, such as video editing, can be supported in preparation of the output.

The following sections detail the process by which the FMCF provides its users functional capability, as well as the current status of the development effort and procedures.

Meeting User Needs

As a service-oriented capability, the FMCF meets user needs by providing the following specific functional benefits to the IC for linguistic translation and processing.

Intrinsic Interoperability. By incorporating standardized capabilities, data models, predictable behavior, scalability, and reliability, the FMCF obviates the need for integration and data transformation to realize a derived service or workflow when two or more FMCF services are integrated by a client application, an orchestrating service, or an enterprise service bus implementation.

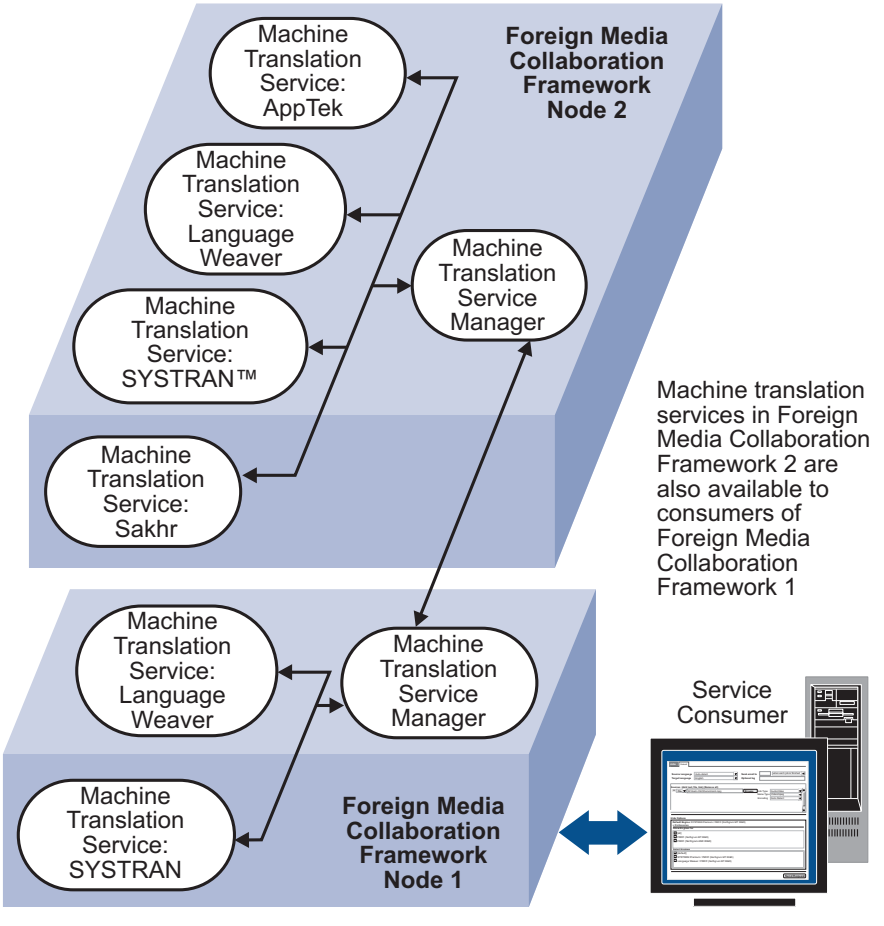


Figure 5. Example of Foreign Media Collaboration Framework distributed processing

Increased Federation. An individual FMCF service can reside anywhere in an enterprise. Standardized interfaces and data models, as well as the FMCF management features, can locate available resources within the framework and facilitate federation. Fault tolerance increases with federation. If a job fails at one engine, it can be sent automatically to an identical engine in another part of an enterprise, as Figure 5 demonstrates. The FMCF also supports limited federation, in which a service hosted at a particular FMCF node is available for use by a subset of FMCF consumers.

Greater Vendor Diversification. The FMCF allows clients to select the "best-of-breed" linguistic engines. That benefit is particularly valuable because languages, dialects, subject domains, and content genre vary from one piece of foreign media content to another. The FMCF abstracts all such decisions and supplies the best choice, given the linguistic criteria set by users at run time.

Better Alignment to Business and Technology Domains. The service contracts for each service in the FMCF are defined to govern a primitive linguistic function, such as ASR, MT, or OCR. This form of governance allows easy integration of each primitive linguistic service by higher processing layers, such as orchestrating services, enterprise service

buses, or applications to mirror an organization's linguistic-based business processes. As business processes change or new ones are added, the primitive services can be recombined to map to the evolving business workflows.

Higher Return on Investment. The FMCF allows Northrop Grumman Information Systems to "integrate once, use anywhere." As new or improved linguistic technologies are added, FMCF users see immediate benefits with no modification required on their part. When we integrate a new linguistic engine, most of the FMCF code base is reused, including all of its management functions. Only the integration glue code, which couples the engine to the service contract, is created anew.

In addition, when an organization purchases a vendor linguistic solution that already exists in the FMCF at another organization (a common practice), the identical FMCF service can wrap the vendor solution and publicize the additional instance of the solution throughout the entire enterprise. The FMCF locates and manages the added service. Each new service added by any member instantly expands the total processing capability of the entire enterprise.

Improved Organizational Agility. The linguistic technology domain is agile in the use of technologies to solve linguistic problems, such as the immediate need for improved linguistic technology to support deployment of United Nations peacekeepers. That need could easily be met by the FMCF, which allows rapid insertion of new or enhanced linguistic technology, language pairs, and other capabilities into the infrastructure, as well as rapid integration and composition of linguistic technologies for users.

Reduced Information Technology Burden. The FMCF reduces the redundancy and inefficiency of requiring linguistic applications to integrate directly with vendor-specific linguistic solutions. Using the FMCF to integrate multiple applications with multiple vendor-specific solutions facilitates a corresponding reduction in infrastructure size, governance, and system evolution costs.

Framework Implementation Status

For more than a year, Northrop Grumman Information Systems has worked closely with IC users in a series of spiral development efforts to prototype and refine the capabilities envisioned for the final FMCF. The MT capabilities are complete. OCR capabilities are in the process of integration. The ASR and the audio/video editing capabilities were combined in the Audio/Video Processor feature, which is currently at a mature prototype level of development.

Design Standards and Paradigms

Northrop Grumman has applied numerous design standards and design paradigms while formulating the FMCF and its service components. Those standards and paradigms include displaying engine features and adhering to community standards. They are examined in the following paragraphs. Key SOA terms are defined in a sidebar, "Key Terms for Understanding Foreign Media Collaboration Framework" (page 57).

Display Engine Features. A potential drawback in defining standard interfaces is the possibility of masking engine-specific features. In addition, a standardized service interface may lack support for new features.

Key Terms for Understanding Foreign Media Collaboration Framework

Certain key terms must be clearly defined before the implementation of SOA in the FMCF can be understood:

- *Parallel corpora:* A segment (usually a phrase) of text in one language presented in side-by-side alignment with its translation in another language.
- *NIST/BLEU:* The National Institute of Standards and Technology/Bilingual Evaluation Understudy scoring algorithm, a method for mathematically evaluating the quality of machine translations of text.
- *Session replication:* The copying of session state information from a particular server that is processing a request from a given cluster to other servers in the same cluster. Such copying minimizes data loss and enables continued session processing if the particular server fails. Session replication provides failover (the ability to switch to a standby computer server/system) but requires more computational resources than does a *sticky session.*
- *Sticky session:* Processing in which only a particular server has the necessary data to process the client request. If that server fails, the client application must create a new session and restart processing from the beginning. Sticky sessions do not provide failover, but they require fewer computational resources than do *session replications.*

Strict adherence to a comprehensive design standard throughout the development of linguistic services in the FMCF can significantly mitigate such risks. All interfaces and integration logic provide visibility of all engine-specific features that can affect the product generated by the engine. Such visibility enables FMCF users to leverage all features that are supported by a particular engine.

Some features are supported by an engine directly, while others extend a native FMCF engine capability. For example, if an ASR engine natively supports Moving Picture Experts Group Layer-1 Audio Layer 3 (MP3) audio format, then that engine processes the file without transformation. If an ASR engine supports only the waveform audio format, the FMCF will convert an MP3 file format into the waveform format as a preprocessing step before invoking the ASR engine.

By enforcing that design standard, the FMCF development team has built extensible service interfaces that enable the user to fully configure engine processing parameters and to define the level of autonomy the FMCF may apply. Users may restrict processing to their exact parameter specifications, allow the FMCF limited autonomy, or allow the system to select the optimal parameters based on predefined default configuration sets.

Adhere to Community Standards. The Northrop Grumman Information Systems FMCF development team adheres to industry and government standards for both interoperable protocols and data. Because the user base for the system comprises multiple DoD and IC agencies, special care was taken to follow the Defense Information Systems Agency's Net-Centric Enterprise Services governance for choosing SOA components.

As a Web-service-based framework, the FMCF leverages the widely used XML-based protocol standards, including Web Services Definition Language XML Schema, Simple Object Access Protocol, and Web Services Security. FMCF security is fully compatible

with public-key-infrastructure digital certificate and Secure Sockets Layer technologies that are widely accepted throughout the IC and the DoD. Moreover, FMCF Web services are fully compatible with multilayer security architectures such as those described by Ellinger in a previous *Technology Review Journal* article [3].

Given the heterogeneous nature of information systems and their data, the FMCF development team also focused attention on the interoperable standards for data. The IC and DoD have made significant strides in normalization of metadata standards relative to the widely used Dublin Core–based discovery metadata definitions. Specific standards include IC XML and the DoD's Discovery Metadata Standard. Both standards incorporate the IC's Metadata Standard for Information Security Markings. Those standards allow for the automatic discovery and cataloging of data independent of content or information system. The FMCF complies with those standards to ensure compatibility with tools available throughout the IC and DoD.

For linguistic processing, the FMCF adheres to International Standards Organization (ISO) 639-1 and ISO-639-2 for languages, ISO-639-3 for languages and language variants (or dialects), ISO-15924 for scripts, ISO-3166-1 for digraphs and trigraphs (sequences of two and three characters, respectively, that are interpreted as one character by the programming language) for countries, and other standards for character-encoding support, as well as industry and de facto standards for text documents, images, audio, and video. TMX XML is used to aggregate parallel corpora data (large collections of parallel texts).

Service-Orientation Paradigm

The FMCF service components were developed based on the eight SOA design principles: standardized contracts, loose coupling, abstraction, reusability, composability, autonomy, statelessness, and discoverability [4]. Applying service orientation to linguistic processing can be straightforward for some principles, whereas others require a careful balance. The following paragraphs explain the rationale behind each service orientation principle as it was applied to the FMCF. The SOA design principles are briefly defined in a sidebar, "Service-oriented Design Principles" (page 59).

Standardized Service Contracts. The service contract defines what a service does and affects all aspects of service design. A key goal of service contract standardization is to increase interoperability via standardized data objects and interfaces.

Linguistic-processing systems share similar data models. Linguistic processing tends to have difficulty with coarse-grain-content documents. Accordingly, the FMCF development team defined a standard content schema that can represent any form of document—text, image, or audio/video—enabling the document to be passed easily from one FMCF service to another without conversion. The FMCF ties all linguistic metadata to the same schema representation.

FMCF services can be used for ASR audio or video content, to translate text documents or OCR images, or to extract entities from text. The development team standardized operations so that each service provides a common set of capabilities defined in its service contract. For example, as Figure 6 shows, all MT-related services, whether acting as a resource manager or as an engine performing the actual translation, share the identical service contract.

The product generated by a service can be influenced significantly by the processing options selected before the operation commences. That complexity requires the careful,

Service-oriented Design Principles

Eight design principles are required for service-oriented applications:

- *Standardized contracts:* All like services share identical definitions for data objects and interfaces.
- *Loose coupling:* When the internal service logic and the service contract are independent, each vendor's engine can be wrapped with the identical service contract for the same type (MT, OCR, or ASR) of service.
- *Abstraction:* Each contract must contain only the essential information about a service.
- *Reusability:* Capabilities are separated from business processes to create simple signatures.
- *Composability:* The best of each type of service that is available for a job can be linked for various needs.
- *Autonomy:* Services do not depend on specific hardware or software components.
- *Statelessness:* The principle of minimizing state maintenance for service requests thereby enhancing scalability and reusability.
- *Discoverability:* Standardized service contracts are published at each service endpoint site and dynamically checked for availability at run time.

deliberate determination of which vendor options should be implemented and by what method.

The options available to MT, OCR, ASR engines can differ across vendor implementations. The FMCF categorizes each option on each vendor's engine as either required, optional, or extension. These categories are defined as follows:

- *Required.* These attributes must be set by a user to perform the core function.
- *Optional.* Considered relatively common across multiple vendors, these attributes may or may not be set by a service user.
- *Extension.* These attributes may be supported by only one vendor or may be features not initially supported by current-generation engines, but expected to be added later. To remain aligned with the stated design standard of displaying all vendor options that can affect the end product, the service contract is designed to include generic descriptors for discovering and setting extension options. The attributes may or may not be set by a service user.

Loose Service Coupling. Loose coupling between FMCF service contracts and the underlying service logic is critical to enabling service contract standardization for each type of service. For example, no matter which vendor supplies the OCR engine, the same service contract is used to specify its functionality. Loose service coupling makes such flexibility possible. Tight coupling exists only between the service logic and vendor engine API, as well as between the service user and the service contract.

Linguistic technologies evolve rapidly, so new or upgraded engines are added frequently. Loose coupling standards allow additions without breaking existing services. If advances in technology force a modification of the service contract, the change is limited to one type of engine, such as the MT or OCR. That feature limits the number of modifications to dependent services.

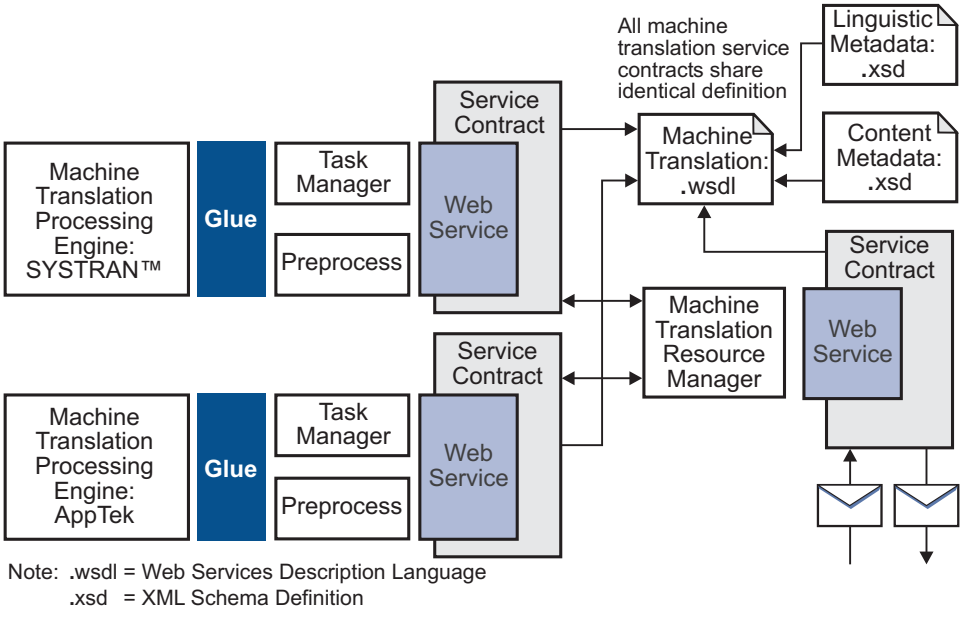


Figure 6. Standardized service contracts for machine translation services

FMCF developers avoid coupling external applications to a service implementation or a specific vendor's engine. For example, a particular application may always assume that the Russian-to-English engine will be provided by one specific vendor and that the vendor's engine includes a medicine domain dictionary. In such a case, the user application cannot simply assume that the dictionary is supported, because at some point the specific vendor's engine may be supplanted by a different engine. The client should retrieve all dictionaries that are available to the user for Russian-to-English and choose the medicine subject domain dictionary, if it is included in the available choices, or the most closely related dictionary in the absence of a medicine subject domain dictionary.

Service Abstraction. The FMCF was built to handle two types of users, each requiring a different level of detail about the capabilities of each service. Most users prefer a highly automated process. They know the content to be processed, the data format, language, dialect, genre, subject domain, and linguistic operation to be performed, but they may not know the best service options. For those users, the linguistic processing services are configured in the FMCF by the user's application to locate and choose the best engine and the right combination of engine options.

Other users prefer to select engine extension options to meet their translation requirements. Those users most commonly employ an application in which a person can provide direct feedback to adjust and control the quality of the linguistic service product. The extension options for each service and related engines are discovered from the service at run time.

For both types of users, the available options for each service are displayed, but the programming logic of each service and associated engines is hidden.

Service Reusability. Because of the constantly evolving nature of linguistic processing, the implementation of each FMCF core service will change often as more capable vendor solutions become available. To accommodate the many add-ons and integrated features

that vendors attach to their products, however, the FMCF has incorporated simple signatures identifying the core capabilities of each service type. Simple signatures allow the service contract and the service itself to be agnostic to specific business processes. For example, MT services perform translation only, and ASR services transcription only. Add-on features that may address specific business procedures are handled after this basic association is made.

If an add-on feature is capable on its own or falls within another FMCF service category, then that add-on feature will be wrapped as separate service. For example, ASR products commonly support MT in a single solution. Ideally, the FMCF does not provide the integrated combination; rather, it provides the necessary primitive services (ASR and MT) instead. An application must then compose the best-of-breed combination of services.

Service Composability. Composability is closely related to the principle of reusability. Initially, the FMCF provided only basic entity services for MT and ASR, but great benefits have been achieved using only a simple service composition. For instance, human translators processing Chinese technical journals previously averaged eight pages daily. The DLIPS application can OCR and MT 300 pages in an hour, albeit with less satisfactory results. When the DLIPS-composed services are used for preprocessing documents for human translators, their efficiency is increased by a factor of three for highly complicated material and a factor of seven for simpler material.

ASR capabilities are currently being added to the FMCF in partnership with vendors such as BBN Technologies and Apptek (Applications Technology, Inc.). Future iterations may incorporate composition controllers to perform common combinations, such as OCR and MT, or ASR and MT. To accommodate a wide variety of existing and future processes, the development team has created a generic set of services that can be composed in diverse ways to support current business processes. That set can be expanded to address as-yet undefined processes.

Composing FMCF services across multiple business processes, however, does involve potential risks that can affect all compositions and their users. We have identified the following risks and associated mitigation methods:

- *Single points of failure.* The FMCF distributes services and their management across multiple organizations in a given enterprise to split the cost of linguistic processing solutions. That federation also provides fault tolerance. If a user knows the end-point site or host FMCF node of at least one active service in the framework and creates a job, then the FMCF will find an implementation within the enterprise that is available to do the work. Nevertheless, this feature does have limits. Some vendor solutions, especially ASRs, are so expensive that they may exist in only one place and may go offline because of a component failure or a quality-of-service constraint. In that case, the only fallback position for the FMCF is to warn the user of a delay and queue the job until the single service becomes available.
- *Performance bottlenecks.* Linguistic processing can be inherently slow, owing to the complexity of the work and the large-scale granularity of the content. To minimize delays, the FMCF implements a priority scheme, ranging from low priority to near-real-time immediate priority. Each FMCF user will have a maximum priority setting, managed by the host organization. That configuration prevents latency-tolerant users—those who specify overnight batch processing, for example—from monopolizing resources needed by users with near-real-time priorities.When vendor solutions permit, a lower priority task could be paused midstream to allow a higher priority task to take over.

- *Service contract rigidity.* If an FMCF service is a member of many service compositions, it is more difficult to modify the contract. Significant risk and cost are associated with potentially breaking all of the applications that use those compositions. To alleviate some risk, FMCF service contracts are defined to support built-in extension options that allow unforeseen features to be included when new vendor solutions are integrated. When a contract is modified to include technological advances, only one service type (MT, OCR, or ASR) should be affected, thus limiting the propagation of changes needed and related effects on users. Further, multiple service contract versions can be supported, so migration to the new service can feasibly be scheduled.

Service Autonomy. In a perfect deployment, FMCF services would have pure autonomy, meaning no shared dependency on hardware and software components. In that state, the services would provide the most predictable and reliable performance. Unfortunately, pure autonomy is impossible, given the nature of the framework. FMCF services will always be tied to their implementation and to the state-of-the-art linguistic hardware and software engines produced by third-party vendors. Further, each vendor tool will have its own limitations, which will influence autonomy.

At a minimum, the services themselves are primitive. Moreover, the service contracts for each service type do not overlap, although FMCF capabilities do demonstrate mixed autonomy. Management capabilities across each service type (MT, OCR, ASR) have shared autonomy. They share databases for resource and configuration management, job queuing, and authorization. Otherwise, the management capabilities are independent, whereas the capabilities related to the linguistic processing operations themselves are linked closely with the vendor engines.

Even the autonomy of vendor engines and the FMCF services that wrap them depends heavily on the deployment at individual organizations. One significant possibility: some organizations may use the engine tied to the FMCF, but operate through separate vendor-provided client interfaces, tools, APIs, and add-ons. Although discouraged, that approach will likely be encountered, given the expense of the high-end linguistic processing engines. It simply may be too costly to support FMCF and legacy linguistic applications with separate engines.

Of course, if that approach is chosen and the FMCF cannot reliably manage the service resources, the result may be unpredictable, perhaps unacceptable, response times. To mitigate such problems, the FMCF supports the marking of individual dual-use services as capable of supporting only medium- to low-priority service requests, as opposed to high-priority or near-real-time requests.

Service Statelessness. In the FMCF, most linguistic processing functions tend to work against text. Thus, a large amount of content is processed in a single step. Depending on the size of the task and processing cycles needed, time to completion can range from seconds to hours. Accordingly, a significant portion of the services must maintain state across multiple service requests. However, the FMCF does defer the storage of state information to back-end storage devices, such as databases and file systems.

That approach allows FMCF services to be clustered without requiring session replication or sticky sessions, since state information in the back-end storage can be accessed by any server in the cluster. Any of the clustered members can handle client requests for status or products after a job is started. Further, the FMCF gains some measure of fault tolerance, since the execution of a linguistic-processing job, previously assigned to a

service that went offline, can be assumed by identical service implementations with only a small latency penalty to detect the failure.

Service Discoverability. The first generation of Web services included discovery—specifically, Universal Description, Discovery, and Integration (UDDI)—as a core component, but that service has not been widely implemented. The FMCF does not provide its own UDDI service. Nevertheless, if an appropriate UDDI service becomes available in the DoD or the IC, the FMCF services will be registered. In the meantime, standardized service contracts can be published at each site that hosts an FMCF node.

Quality-of-service information will depend on each hosting organization, but FMCF resource management functions can compensate for intermittent service uptimes and downtimes by dynamically determining resource availability at run time and working around quality-of-service issues across organizations that compose the FMCF enterprise. That built-in management feature reduces the logic required by the clients to perform the discovery function. In addition to the service contracts, each service provides capabilities to dynamically determine the options available for each type of service, including the extended options that may be available from a single vendor solution.

Types of Services Provided

The end product of the FMCF effort is a collection of linguistic-processing services—including MT, OCR, parallel corpora, and document management—that can be reused and composed to satisfy the needs of a variety of business processes. Depending on the input format—e.g., image, text file, audio file, video file—as well as the desired output, the process may require the use of some or all of those services. Of particular importance is the document management service, which maintains continuity in the output of each service and offers options for the output of the entire business process.

Machine Translation Service. The MT service provides the operations necessary to perform a single-language translation of textual content. The textual content can be as simple as a single word or as complex as an entire Microsoft® Office Word–compatible document. The MT service includes document parsers to handle content types not natively supported by some vendor MT solutions, as well as to maintain the formatting and images contained in the original document.

Northrop Grumman Information Systems is working closely with certain vendors to incorporate additional language pairs and new capabilities currently in development. Table 1 lists the currently supported foreign languages; in all cases, the other half of the supported *language pair* is English. Most vendors offer language support for bidirectional translation (foreign language to English, as well as English to foreign language), but a few support only monodirectional (foreign language to English) translation.

In addition to the translation operations themselves, the FMCF supplies resource management and locator functions, as well as the decision logic to choose the best-of-breed MT engine. Currently, the FMCF supports only single-language translations. If a text source contains multiple languages, an orchestrating user must identify and provide those specific text sections separately to the service to be translated.

Optical Character Recognition Service. The OCR service provides operations necessary to extract text displayed within an image. Image format conversion is built in to help vendor OCR implementations that do not support all types of images. A single image or a group of images can be processed into a single document. The service also enables

Table 1. Language support provided by initial Foreign Media Collaboration Framework deployment

Language	Machine Translation				Optical Character Recognition			Automatic Speech Recognition: BBN Technologies AMC[a]
	SYSTRAN™	Sakhr	Language Weaver	AppTek	ABBYY® FineReader®	Sakhr Reader	Readiris™	
Albanian	○				△			
Arabic	◇	◇	◇	◇		△		△
Bahasa				◇	△			
Chinese	◇		◇	◇	△		△	△
Dutch	◇		◇	◇	△			
French	◇		◇	◇	△			
German	◇		◇	◇	△			
Hindi			◇					
Italian	◇			◇	△			
Japanese	◇			◇	△		△	
Korean	◇		◇	◇	△		△	
Persian	○		◇	◇		△		
Polish	○			◇	△			
Portuguese	◇		◇	◇	△			
Somali			◇					
Serbo-Croatian	◇				△			
Romanian			◇					
Russian	◇		◇	◇	△			
Spanish	◇		◇	◇	△			
Swedish	◇		◇		△			
Tagalog				◇	△			
Turkish				◇	△			
Ukrainian	○			◇	△			
Urdu	○			◇		△		

○ Monodirectional ◇ Bidirectional △ Supported [a] Audio Monitoring Component

named areas of an image to be defined. An area can be flagged to be either processed or ignored, as appropriate. That capability allows the user to identify pieces of metadata, such as title and author, even before the content is in textual form. As it does with the MT services, the FMCF provides resource management and can choose the best OCR engine for the job, given the image linguistic information.

Parallel Corpora. A *parallel corpora* pair is a segment (usually a phrase) of text in one language presented in side-by-side alignment with its translation in another language. The FMCF parallel corpora service presents text in a source language, dialect, and genre paired with its translation in another language, dialect, and genre, and maintains metadata about

each segment. Parallel corpora pairs are used by researchers in the linguistic community, by translators as a phrase dictionary to assist translation, and by MT vendors as a source of ground-truth data for statistical MT and translation memories, as well as performance analysis. Through the use of that service, the FMCF can create parallel corpora with linguistic, bibliographic, and quality metadata; update and define versions of existing parallel corpora; and discover, import, and export the parallel corpora.

Document Management. The document-processing management service provides a mixture of document-related capabilities grouped as a single service. The integrated service supplies capabilities required for postprocessing cleanup and format simplification of OCR-generated documents to prepare those documents for use by MT engines. Absent the cleanup, the resulting documents cannot be further processed through MT or product generation without degrading the source text. The simplification capability also supports document composition to combine several single-page documents intelligently into a multipage document that maintains format and page breaks.

The service can generate Adobe® Portable Document Format (PDF) files from source images or formatted text documents. The FMCF can even take two text documents or scanned image sets that contain parallel source and translated text and combine the pair in a single PDF, displayed side-by-side. Other capabilities include the detection of character encoding in HyperText Markup Language (HTML) or plain text documents and language detection.

Existing Products

Foreign Media Collaboration Framework Portal. Currently in testing is the FMCF Portal, which leverages Microsoft® Office SharePoint® Server 2007 and the off-the-shelf capabilities of SharePoint to manage artifacts and products. The FMCF Portal is a central location at which users can download the Translator Workbench software, manage document workflows, access MT capabilities, and publish resulting products. Future versions of the FMCF Portal will incorporate ASR, audio/video editing, and entity extraction. Figure 7 shows the current architecture of the FMCF Portal.

Digital Library Input Processing System. The Northrop Grumman–developed DLIPS application currently supports a broad range of foreign media processing capabilities using hard- or soft-copy source documents. The core capabilities include OCR in several languages, translation using several MT engines, and the generation of products in several formats. A key feature of the DLIPS application is flexible metadata tagging. Users can identify the metadata they wish to capture by creating and maintaining their own set of schemata using DLIPS tools. For most clients, the tagging is used to provide discovery, bibliographic, and security classification metadata.

Translator Workbench. The Northrop Grumman–developed Translator Workbench is used by human translators to assist and simplify the translation process. The Translator Workbench

- Maintains and manages document layout and images
- Enables individual source and translation pairs of text to be extracted as parallel corpora

The human translator can focus on only the text to be translated for each document section, paragraph, or caption since document formatting is programmatically performed. He or she can also perform quality check, realignment, and further refinement of the parallel corpora, as well as add associated discovery metadata.

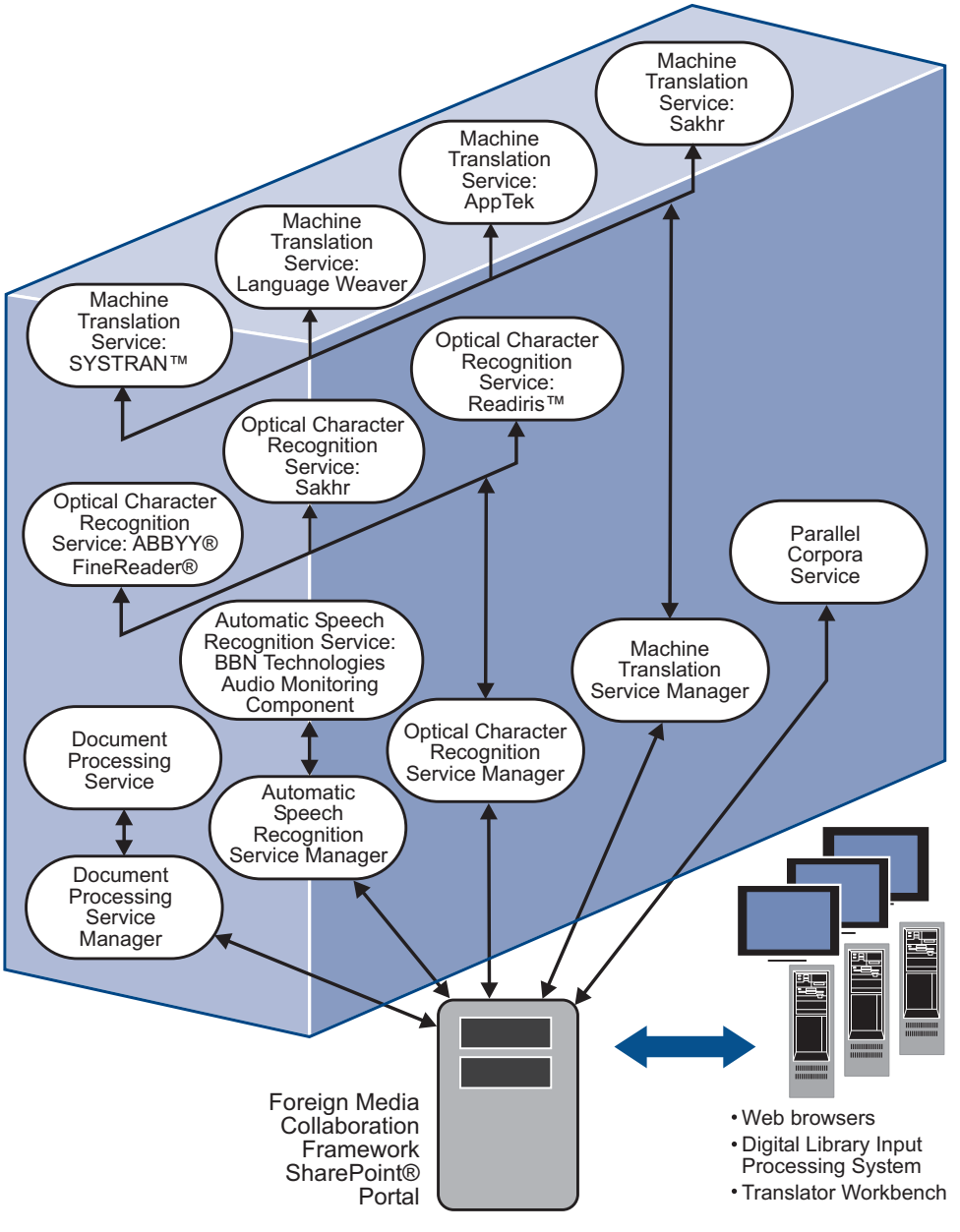


Figure 7. Architecture of Foreign Media Collaboration Framework Portal

Web Machine Translation. On the surface, Web MT is a browser-based machine translation client that can be as simple to use as Google Translator or Microsoft® Windows Live™ Translation services. However, the Northrop Grumman Web MT application translates more than just HTML and text. Any document in a Microsoft Word–compatible format, including extremely large documents (800+ pages), can be sent for translation from the Web-based client. Web MT also allows users to select from the complete set of options offered by any available vendor engine.

Parallel Corpora Manager. The Parallel Corpora Manager is a suite of tools that support the creation, import, and export of parallel corpora data. Users can assign metadata that comprises linguistic and bibliographic information to an individual record, group of records, or an entire imported file. The tool accepts and provides parallel corpora in TMX format, as well as a parallel corpora XML format defined in collaboration with the National Virtual Translation Center.

Machine Translation Evaluation Center. The FMCF MT Evaluation Center enables users to evaluate MT engines via the NIST/Bilingual Evaluation Understudy (BLEU) scoring algorithms, the same algorithms used during the NIST MT evaluations. The Center can be integrated with DLIPS as an add-on feature or a stand-alone test kit. The NIST/BLEU algorithms generate a score that is based on a statistical comparison of a text segment produced via an MT translation and a text segment considered to be a ground-truth translation. Though lacking a linguist's subjective rating of translation quality, the score can provide a coarse-grain view of the translation quality.

Products under Development

Automatic Speech Recognition. An ASR service is currently being added to the FMCF. Leveraging capabilities from vendors including BBN and Apptek, the ASR service will abstract state-of-the-art ASR tools for generating segmented, time-stamped transcripts of audio and video content.

When composed with other services and applications, the resulting transcription products can be translated by the MT service or edited by a human translator, packaged in a Moving Picture Experts Group Layer-4 (MPEG-4) container as timed text (subtitles), overlaid on video content using the Audio/Video Processor editing service, added to discovery metadata for the content, or used as a stand-alone product. The BBN Broadcast Monitoring System™ Audio Monitoring Component v3.0 engines provide the initial capabilities in Modern Standard Arabic and Chinese and tie the text to specific frames of the video content as subtitles. The service contract can abstract other ASR implementations that can be integrated under later efforts.

Audio/Video Editing. An audio/video editing and processing service capability is currently in development. Initially, it will provide ASR preprocessing format conversions between audio-to-audio and video-to-video. The service will also support the generation of audio and video products after ASR, including combining video with text overlays of transcribed or translated speech as a single video product, or packaging a video with timed text in a single MPEG-4 container file.

Entity Extraction. Currently planned as a phase II extension of the Northrop Grumman Audio and Video Processing Independent Research and Development (IR&D) project, the Entity Extraction service will abstract state-of-the-art tools to identify proper nouns, such as people, organizations, geographic locations, currencies, and equipment within textual

content. Initially, the service will plug in the Microsoft™ FAST Enterprise Search Platform™ (FAST ESP™) to meet the defined service contract. User applications can insert the entities into discovery metadata, then combine or detect relationships between entities for storage and link analysis. The FMCF development team is also implementing related capabilities from Basis Technologies and evaluating other vendors for potential incorporation into this task.

Linguistic Applications. The predecessors of the FMCF (DLIPS and the applications built on it) are currently installed and operating in production environments in NASIC and ten other national and international organizations. Although the applications themselves are highly successful, the predecessor Web services they leverage were built initially with a single application (DLIPS) in mind; they were not meant to support enterprise applications. In contrast, using the lessons learned from DLIPS, the Northrop Grumman Information Systems FMCF development team applied service-oriented principles from the beginning to provide enterprise-level linguistic processing services.

Conclusions and Future Plans

Many organizations require state-of-the-art linguistic processing engines to perform their missions, but they are finding it more and more difficult to take on the entire cost burden. As mission needs and corresponding business processes evolve to require more linguistic technologies, the burden will increase correspondingly.

From its inception, NASIC has played a key role in linguistic processing, and Northrop Grumman Information Systems has been a steadfast partner. NASIC strongly views service orientation as the solution to the significant cost challenges ahead for both its own organization and other IC and DoD organizations. The result of that vision is the service-oriented FMCF for foreign media linguistic processing.

NASIC will be the first to deploy the FMCF and demonstrate its linguistic capabilities to the entire IC and DoD. NASIC will also form the initial project management office and membership of the governance board for the FMCF implementation. The board is expected to include at least one member from each organization deploying a node.

IC and DoD organizations with no inherent linguistic tools are now offered secure access to NASIC's entire linguistic suite. The next step is to expand further use of the FMCF by other IC and DoD organizations, which can then share their in-house linguistic resources via the FMCF collaboratively. That expansion will enable FMCF customers to achieve the overarching goal of shifting the focus of linguistic processing from budget constraints to collaborative mission execution.

References

1. U.S. Department of Commerce, National Institute of Standards and Technology, *NIST 2006 Machine Translation Evaluation Official Results, MT-06*, Gaithersburg, Maryland, November 1, 2006, *http://www.nist.gov/speech/tests/mt/2006/doc/mt06eval_official_results.html*. Accessed August 10, 2009.
2. T. Erl, *SOA Principles of Service Design*, Prentice Hall Service-oriented Computing Series from Thomas Erl, Prentice Hall PTR, Upper Saddle River, New Jersey, 2007.

3. R.S. Ellinger, "Service-oriented Architecture and User Interface Services: The Challenge of Building a User Interface in Services," *Technology Review Journal*, Vol. 15, No. 1, Spring/Summer 2007, pp. 43–60, *http://www.is.northropgrumman.com/about/ngtr_journal/assets/TRJ-2007/SS/07SS_Ellinger.pdf*. Accessed August 10, 2009.
4. R. High, Jr., S. Kinder, and S. Graham, *IBM's SOA Foundation: An Architectural Introduction and Overview*, Version 1.0, IBM Corporation, Armonk, New York, November 2005, *http://download.boulder.ibm.com/ibmdl/pub/software/dw/webservices/ws-soa-whitepaper.pdf*. Accessed August 10, 2009.

Acknowledgments

The authors are indebted to Douglas S. Barnhart, formerly of Northrop Grumman Information Systems, for his outstanding contributions to the design, planning, and early development of the FMCF concept. Thanks also to David M. Barber and Charles D. Simmons, National Air and Space Intelligence Center, for their support and encouragement during the performance of the work discussed here. Additional thanks are due to Northrop Grumman Information Systems staff members Roy E. Kimbrell, Dario DeAngelis, and Kevin T. Sculley for making the IR&D initiatives possible, as well as to Kimberly W. Fike for exemplary editing assistance.

Author Profiles

Garin R. Clint is the program manager for the Textual Information Systems group at Northrop Grumman Information Systems, Scientific and Technical Intelligence Systems business unit. He has worked in the field of foreign media collaboration throughout his 17 years at Northrop Grumman. During his tenure, Mr. Clint has designed, developed, and implemented multiple complex foreign media collaboration systems for a wide variety of customers in both the Department of Defense and the intelligence community. Those systems are currently deployed worldwide and in use by both U.S. and coalition forces. After focusing on the applied sciences for the major part of his career, Mr. Clint has recently begun working in the area of basic research for Northrop Grumman. He holds both a BS in management information systems and an MBA from Wright State University.

garin.clint@ngc.com

Richard J. Thomas is currently serving as principle software engineer for the Textual Information Systems at Northrop Grumman Information Systems, Scientific and Technical Intelligence Systems business unit. He provides technical leadership for the FMCF tool suite. He previously served as senior software engineer leading the Parallel Corpora project. A holder of five patents, Mr. Thomas has served as chair of the Institute of Electrical and Electronics Engineers, Dayton Section, and is the recipient of numerous awards. He holds BS and MS degrees in electrical engineering with specialization in control engineering, both from The Ohio State University.

richard.j.thomas@ngc.com.