(U//FOUO) Catalyst Entity Extraction and Disambiguation Study Final Report (U)

Prepared for:
IARPA/RDEC

Prepared by:
Dr. A. Joseph Rockmore/Cyladian Technology Consulting

21 June 2008

(U//FOUO) Executive Summary (U)

(U) Catalyst, a component of DDNI/A's Analytical Transformation Program, will process unstructured, semistructured, and structured data to produce a knowledge base of ***entities*** (people, organizations, places, events, …) with associated attributes and the relationships among them. It will perform functions such as *entity extraction*, *relationship extraction*, *semantic integration*, *persistent storage of entities*, *disambiguation*, and related functions (these are defined in the body of the report). The objective of this study is to assess the state-of-the-art and state-of-the-practice in these areas.

(U) The approach to the study was to survey relevant commercial and open source products, and to identify and describe relevant government systems. (Work in academia was postponed.) Over 230 products and over 20 government systems were identified and analyzed. The conclusions and recommendations are summarized below.

(U) In the commercial and open source worlds, there is significant activity to deliver functions needed by Catalyst. By function, the conclusions of the analysis are:

- *Entity Extraction* (14 open source/50 commercial products) – There are many products, with Aerotext, Inxight, and NetOwl best known. There have been no significant breakthroughs in capability or performance recently. Some vendors claim precision and recall over 90%, but with varying scoring definitions.
- *Relationship Extraction* (2/24) – Nearly all are an adjunct to Entity Extraction. Again, Aerotext, Inxight, and NetOwl are best known. Performance is typically poor, with precision and recall as low as 60%. Customization for specific domains is difficult, time-consuming, and with unknown performance.
- *Semantic Integration* (6/15) – There are many tools that do integration (mostly from the database world), but few are focused on semantic integration. There does not seem to be recognition by vendors of the need for this function.
- *Entity Disambiguation* (2/8) – Relatively few products perform this function. Mature and proven ones are often only in a non-government context. There is a lack of comparative measures of performance.
- *Knowledge Base* (17/17) – There are quite a few available, most specialized for storage and retrieval of semantic data, and most based on Semantic Web standards such as RDF and OWL. The performance of the tools, such as load time and query time for some standard queries, is often difficult to determine.
- *Visualization* (18/36) – Many perform some kind of analysis (e.g., link analysis) and visualize the results. Not mature (as applied to entity data).
- *Query* (18/36) – Most products are connected to some knowledge base approach, which defines the query approach.
- *Analysis* (4/40) – Few complex analyses are done on entity data.
- *Ontology/Data Model* (26/15) – The semantic community has focused on data modeling and tools for building and managing ontologies to date, so there are some capable, mature products.
- *Reference Data* (7/1) – There are many sources of reference data that were not discovered during the study, due to resource constraints. The ones described are government databases that are openly available at no cost.

(U//FOUO) In the IC, there are a small number of relevant programs, with some showing significant capability to deliver functions needed by Catalyst. It is interesting to note that nearly every major IC Agency has recognized the need for Catalyst functionality and has allocated resources to developing capabilities. Some observations are:

- Entity extraction, and to a lesser degree relationship extraction, is a well-funded, active area of development across the IC. The persistent issues with these programs seem to be (1) quality of output, (2) throughput, and (3) difficulty of development. Many programs have determined that to get high quality extraction, the products must be tuned to the particular types of documents and domain of analysis; this is often a long and complex process.
- Once information is extracted from documents or other data, some systems are a stateless service that extracts information and provides the extraction back to the requester, while others (most of them) persist the data. The extracted information is stored according to a data model that captures the salient information in the domain. Very few organizations have gone as far as an RDF triple store with OWL support; more often, the results are stored in a relational database management system. Few systems have scaled up to realistic sizes of entities for real-world intelligence problems.
- Semantic harmonization and integration has been attacked in several of the IC systems, but mostly in an ad-hoc manner; there are few cases of mapping from schemas or ontologies into a common ontology. Surprisingly little has been done in integration; in most cases all property values are kept.
- Disambiguation is understood to be important in most of the entity integration systems. Some IC systems have written custom code, while others have used commercial products. Only one system has delved deeply into this function.
- No IC system has yet integrated existing entity knowledge bases.

(U//FOUO) What emerged from these analyses is a nascent technical area that has great promise, but is still in its infancy. The study recommendations are:

- Continue to perform the kinds of tracking and evaluations that have been done herein, to provide additional reference data and to see if the observations and conclusions of this report remain in effect.
- Where the IC needs to represent common entity classes and common attributes and properties, appropriate groups should be empowered to develop standardize languages and ontologies. Processes should be stood up to develop and manage core ontologies, and all local ontologies should be rooted in the core ontologies.
- Any eventual Catalyst implementation will have to deal with some serious security concerns. Thus a recommendation is to elucidate and analyze the security requirements that are unique to Catalyst.
- A software architecture for Catalyst-like capabilities across the IC should be developed and services of common concern stood up where possible, in a Service Oriented Architecture.

(U) Table of Contents

1. (U) Introduction (U)

(U) The objective of this study is to assess the state-of-the-art and state-of-the-practice in entity extraction and disambiguation in the academic, the government, and the commercial worlds for potential use by the Catalyst Program, a component of DDNI/A's Analytical Transformation (AT) Program. The AT Program is being executed by a variety of Executive Agents, managed by the DNI/CIO. We interpret this purpose to include all ancillary functions needed to develop an end-to-end system that takes as input unstructured data (primarily free text, with or without document-level metadata) and results in a knowledge base of entities (people, organizations, places, events, etc.) with attributes of these entities and the relationships among these entities. Thus, in addition to entity extraction and disambiguation, an eventual Catalyst capability will need functions such as relationship extraction, semantic integration, persistent storage of entities, and others to provide end-to-end functionality. Note that we are not defining what constitutes a Catalyst *system*, but rather what capabilities need to be performed by some processing component to result in the kinds of outputs envisioned for Catalyst, which are sets of semantically aligned, integrated, disambiguated entities of interest for a problem area.

(U) The study was performed by executing the following five tasks:

Task 1—Define the scope and terms of reference for the study. Define the common parameters on which approaches to the problem will be described and the performance metrics for assessment in Tasks 2-4.

Task 2—Perform research on approaches to semantic integration and disambiguation in the academic world, focusing on research in universities and institutions that address advanced techniques for higher performance than is currently available in commercial and government approaches. Include, where possible, assessment of the maturity of the approaches, and the performance (time to execute) of implementations for scaling to realistic data volumes.

Task 3—Perform research on approaches to semantic integration and disambiguation in the commercial world, focusing on existing products and, where the information is available, on the direction the leaders are taking in their product development. Specifically assess their ability to scale to realistic data volumes by assessing benchmarks (if available) for speed of execution as a function of data size and complexity. Include assessment of performance (percentage of correct and incorrect associations) as a function of the underlying data characteristics.

Task 4—Perform research on approaches to semantic integration and disambiguation in the government world, focusing on existing systems and under-development systems, describing approaches and performance, if available. Since it is expected that many government systems will be based on commercial products, describe the process that was used to arrive at the product used and evaluate the lessons learned in using the product(s). If the system is not based on commercial products, describe the process that was used to arrive at the approach, specifically if commercial products were assessed and the reason why they were rejected.

Task 5—Document the study, primarily the results of the research on approaches, but also trends and recommendations for how to take advantage of the study results for the Analytic Transformation Program.

(U) The remainder of this report contains 5 sections. In Section 2 we provide the **Processing Context** that defines what functions Catalyst will perform, with supporting details in Appendices A (Terminology) and B (Detailed Description of Functionality). Second, the **Study Approach** described how the study described in this report was done. Third, the results of the study of **Commercial Products for Catalyst** is presented in Section 4, with supporting data in Appendix C. Then, the results of the study of **Government Systems for Catalyst** is presented in Section 5, with supporting data in Appendix D. Last, **Conclusions and Recommendations** are presented in Section 6.

(U) Please send any comments and/or suggestions to Joe Rockmore, 650/614-3791, or rockmore@cyladian.com (Internet) or rockanj@cia.ic.gov (JWICS).

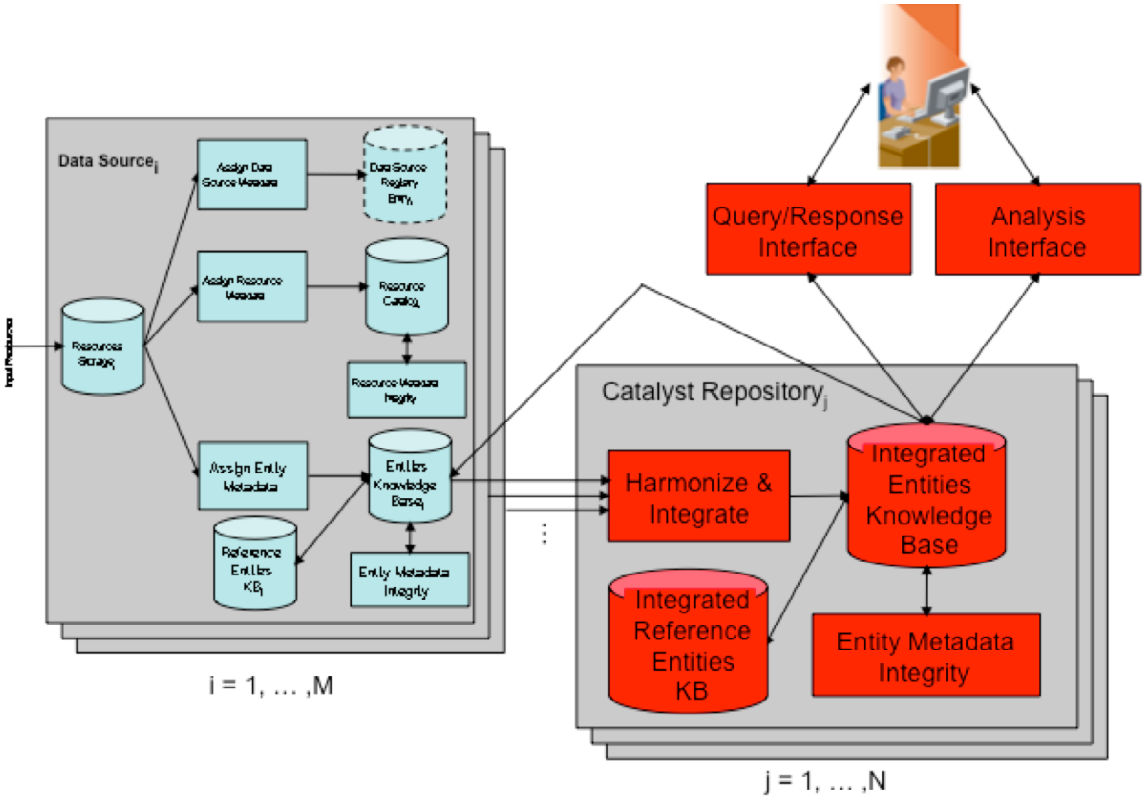
(U) The contributions of Brand K. Niemann and Kelly Wise of SAIC to the commercial product data collection and analysis are acknowledged and appreciated.

2. (U) Processing Context (U)

(U) In order to make sense of the various advanced processing approaches, products, and programs that are in some way relevant to Catalyst, this section summarizes a generic set of functions for intelligence processing to provide the context for the study. A more detailed description is found in Appendix B. The terms used in this report are described in Appendix A.

(U) We assume that this generic processing starts with unstructured and semi-structured data, such as documents, images, videos, audios, signals, measurements, etc., as well as structured data, that are collected from a wide variety of sources by a variety of methods. We use the term *resource* to include all of these input data types. The objective of Catalyst's advanced intelligence processing is to identify the ***entities***—people, places, organizations, events, etc.—in the resources and what the resource says about the entities (the attributes of entities and the relationships among them). This information is made available to users (generally, intelligence analysts) so they can retrieve information about these entities and detect patterns of interest to their analysis mission.

(U) At a high level, there are three steps to the kind of intelligence processing related to Catalyst: (1) *describing resources and making them capable of being processed,* (2) *semantically integrating entities of interest to a specific task* (including disambiguation of these entities), and (3) *processing the entities to produce some conclusion of interest*. The figure below summarizes the three steps of intelligence processing. Each step is expanded upon below.



(U) Describing resources for processing

(U) In order to process the entities in resources, they need to be explicit in a structured form. Some resources are naturally structured (data that is typically in a relational database management system, for example). Other resources are either unstructured or semi-structured[1], such as a document, an image, a video, or a signal. In order to make resources capable of being processed by an eventual Catalyst system, structure in the form of *metadata* must to be added to them. In general, metadata that describes intelligence resources falls into three categories: *descriptive metadata*, *structural metadata*, and *content metadata*. *Descriptive metadata* provides information about the resource as a whole, such as title, authoring agency, security classification, and date of publication. *Structural metadata* describes how a resource is laid out for rendering. *Content metadata* describes what the resource "is about;" it can relate to the resource as a whole, such as the topic or geographic area that the resource is about, or it can relate to the details inside the resource, such as the specific entities mentioned in the resource and what the resource says about these entities (that is, attributes of or relationships among the entities). In this study we are only concerned with content metadata, since that is what Catalyst will operate upon[2].

(U) With reference to the Figure above, an organization will stand up one or more Data Sources, each a collection of resources of a particular type or on a particular topic. Each Data Source receives new resources by some mechanism. The resources typically are stored persistently for retrieval, and three metadata processing steps can be performed[3]. Each Data Source needs metadata to describe the Data Source as a whole, used for *discovery* of Data Sources that may be of use to a particular intelligence tasking. This metadata is provided to a Data Source Registry for search. In addition, all resources are assigned *resource metadata*, which includes all descriptive metadata and that part of the content metadata that is about the resource as a whole. This resource metadata is stored in a resource catalog, so it can be searched and relevant resources retrieved. These processing steps are not the subject of this study.

(U) All resources also can be assigned *entity metadata*; that is, the entities in the resource are identified, delimited, and assigned to a class and, where possible, the attributes and relationships among the entities in the resource are identified. The entity metadata can be stored in an Entity Knowledge Base. The Entity Knowledge Base often includes *reference entities*–representations of well-known and accepted real world entities–that are

---

[1] (U) Unstructured data, such as documents or images, have no inherent structure that describes them, while semi-structured data has an unstructured part–the text of the document or the image–and a structured part that describes the unstructured part, such as the author, title, date of publication, etc.

[2] It is likely that descriptive metadata can help inform the processing of the content metadata, but we do not pursue this idea in this report, since it is a research issue.

[3] Today not many Data Sources perform all three processing steps, but some version of these steps will be necessary for the high quality processing that Catalyst can provide.

not derived from any resource, but input into the Entity Knowledge Base by some other mechanism. There are two major issues regarding entity metadata. The first is *how* it gets assigned, and the second is *what* metadata is assigned. The "how" is accomplished by users, software tools, or a combination of the two. The "what" is defined by each organization and represents the important types (classes) of entities, and the important attributes and relationships of each class, that the organization needs to perform the processing to meet its mission objectives. The classes of entities that are commonly extracted are *person*, *place*, *organization*, *event*, or *thing*, while the kinds of attributes and relationships that are extracted tend to be more specific to the organization. Thus the function usually called "entity extraction" actually encompasses entity identification, entity type evaluation, entity attribute extraction, and relationship extraction. Whereas we include this function in this report, we would expect that this processing is done by the Data Source and **not** an eventual Catalyst system.

(U) Semantically integrating entities

(U) Once the Data Sources have been processed to derive their associated Entity Knowledge Bases, Catalyst will provide an integration of these entities for processing across the Data Sources. This is the major *raison-d'être* of Catalyst: no one organization will integrate across Data Sources that span many organizations, but that is precisely what Catalyst will do. The objective of this cross-organization integration is to enable intelligence analysis on "all we know" about entities, which implies integrating the entity data from various Data Sources from various organizations. The means by which such integration is done is to partition the Data Sources and then integrate the Entity Knowledge Bases from each set of Data Sources into a common Integrated Entity Knowledge Base, part of a Catalyst Repository in the Figure. It is expected that there will be more than one such Repository, based on the partitioning of the Data Sources (see Appendix B for more information on the partitioning of Data Sources into Repositories).

(U) As shown on the Figure, the data must be *harmonized* and *integrated* before it is stored in the Integrated Entities Knowledge Base. Harmonization is a step that brings the entities from various Data Sources' Entity Knowledge Bases into common semantics. The entities might not be in common semantics from their Entity Knowledge Bases due to factors such as scaling and unit differences, different levels of granularity, different definition of concepts, etc. Basically, each entity must be mapped from the semantics of its Entity Knowledge Base into the semantics of the Integrated Entities Knowledge Base. Then the attributes and relationships must be integrated in an appropriate way (different for each attribute or property). All of this processing must not destroy the connections back to the original resources (called pedigree), since the original resource will remain the definitive source for information.

(U) An important part of integration of entities is the processing for integrity. There are many kinds of integrity that may be useful, but one of the most important is *disambiguation* (also called co-reference resolution). This processing is to find multiple entities in a knowledge base that actually refer to the same person, place, organization, etc. in the real world, and combine them. This processing is important since the intelligence analysis potentially needs to use *all* that is known about an entity, which

requires combining all that is known in the separate Entities Knowledge Bases into a single entity in the Integrated Entities Knowledge Base.

(U) Processing entities

(U) The Integrated Entities Knowledge Base will be used by intelligence analysts to support their analysis tasking, resulting in higher quality analysis than available today[4]. There could be a query interface that allows analysts to search the Integrated Entities Knowledge Base for entities of interest, and there can be analysis tools, such as visualization or link analysis, that interface to the Integrated Entities Knowledge Base. In both cases the underlying functions to integrate entities and their attributes and relationships will provide better data against which to work, by virtue of the semantic integration and disambiguation of entity data from many Data Sources.

(U) Querying an Integrated Entities Knowledge Base requires capabilities that are specific to entities and their attributes and relationships. It is not the same as, for example, querying a relational database. Thus the query interface will need to provide functionality to allow an analyst to fully explore the set of integrated entities. There will be times when a query will result in a small number of entities, which the user then can simply view, but this will not always be the case (in fact, will usually not be the case). More often, the results of a query will be a large number of entities, and then there needs to be methods and tools to facilitate visualization and analysis of the data. Such visualizations can include timeline displays or geographic displays of the entities, thus helping the analyst understand the set as a whole. Another analyst capability would be successive refinement of queries (currently called "faceted search"), a process that helps an analyst make good queries by providing feedback on the makeup of a set of entities derived from a broad query, so he or she can see explicitly how to refine the query.

(U) One particularly significant analysis that will be done on the entities is the identification and classification of patterns of interest in the data. Patterns are partially (or fully) uninstantiated sets of entities and relationships, and can be models of behavior of interest (like behavior leading up to a terrorist attack on a certain type of asset). Searching for patterns is a very important use of an Integrated Entities Knowledge Base, so tools will need to support expression of patterns, search for patterns, analysis of results, and presentation of results.

(U) Whereas analysis and visualization of the integrated entities is useful, another very important use of the integration process is to provide information back to the Entities Knowledge Bases of the Data Sources. Two kinds of information may be provided that will be of high value: disambiguation results, and enriched attributes and relationships. The first type of information is to inform the Data Source's Entities Knowledge Base that two or more entities that it provided are in fact the same person, place, organization, etc. in the real world, based on the addition information provided by the other Entities

[4] (U) Today analysts have no choice but to read resources, extract entities and their attributes and relationships manually, keep this data in some local form such as a spreadsheet or Analyst Notebook diagram, and manually integrate across differing Data Sources. Furthermore, they have no automated mechanism to share what they learn.

Knowledge Bases. That is, the original Entities Knowledge Base did not contain sufficient information to disambiguate entities, but when this information is combined with information from other Entities Knowledge Bases, and the information is integrated, additional disambiguation decisions can be made, and this provided back to each Entities Knowledge Base. The second type of information can provide the original Data Source's Entities Knowledge Base with attribute and property values that it did not have based on its resources, but that some other Entities Knowledge Bases provided from their resources. Thus the values of attributes and properties can be greater by virtue of integration, and this information provided back to all the Data Sources.

3. (U//FOUO) Study Approach (U)

(U) Commercial products

(U) The study addressed commercial and open source[5] products by identifying a set of products that could be located (within the study resource constraints) and describing them according to:

- COTS or open source
- Name of product
- Name of company that offers the product
- Functionality product offers, one or more of:
    - Entity Extraction
    - Relationship Extraction
    - Semantic Integration
    - Entity Disambiguation
    - Knowledge Base
    - Visualization
    - Query
    - Analysis
    - Ontology/Data Model
    - Reference Data
- The url where the product description may be found
- A one-line description of the product

(U) All the information collected was done by research on the Internet at the company's web site. As such, no company's claims were validated, and all information should be interpreted as derived from marketing material. There was not sufficient time to contact companies for additional information, nor for any reviews by independent third parties to be discovered and digested. In cases where a Government evaluation was done, an attempt was made to validate any information against the Government evaluation.

(U) Once the data was gathered, it was examined to attempt to draw industry-wide conclusions about the information. That is, where the data was available to collect and analyze, we attempted to discern industry trends and directions related to specific features and performance of the products.

(U) Thus the data in the section on commercial products and its accompanying appendix can be viewed as reference data, as well as data for analysis of the industry. Recommendations for further work in this area are in the final section of this report.

[5] (U) We treated open source products as commercial products with no cost, and sometimes no organization identified that is responsible for maintenance and enhancement.

(U) Government systems

(U//FOUO) Systems being developed or in use in the Intelligence Community were identified and described. By the very nature of the community, it was not possible to identify all such programs; the focus tended to be on the national intelligence agencies around the Washington, DC area and selected other agencies that were known to the report's authors. There is little doubt that there are other systems being developed within the Intelligence Community that might have been included in this report, but were not. Within the time and resource constraints of the study, systems that could be identified and described were, and any conclusions presented herein are limited to those systems. Future work may include other systems and this might modify any conclusions that were drawn.

(U) Each identified government system was described according to:

- Program Name
- Sponsoring organization
- Performing contractor(s)
- Government Point of Contact Phone Number & E-mail Address
- Contractor Point of Contact Phone Number & E-mail Address
- Abstract description
- Intended users
- Catalyst functionality included (according to the same functions as commercial products)
- Sources of input data
- Scale of current implementation
- Status of system
- Where deployed
- COTS/OS/GOTS used
- Size of development effort
- User experiences
- Plans for continued development
- Lessons learned
- Details

(U//FOUO) In some cases not all of the above information was available to be captured. No classified information is included in this report, so some descriptions are necessarily incomplete. In some cases the government and contractor points-of-contact were asked to provide the information, while in most cases the study authors visited the organization developing the system and interviewed the points-of-contact for the information. When the latter was done, the points-of-contact were given an opportunity to modify and enhance the draft descriptions written by the study authors. Due to the brevity of the descriptions, there was often additional information that the points-of-contact offered that did not make it into the report. Specifically, there was information on issues such as security architecture, sharing of data, how to deal with legal issues, etc. that were not captured. The focus was on the functionality or the systems related to Catalyst.

(U//FOUO) As with the commercial products, we tried to figure out Intelligence Community trends and directions related to an eventual Catalyst system.

(U) Recommendations for further work in this area are in the final section of this report.

(U) Academic work

(U) Due to resource constraints, the Government directed that academic work be done at a later date, so no information on Catalyst-related research is presented in this report.

4. (U) Commercial Products for Catalyst (U)

(U) In Appendix C we present 232 commercial and open source products. The number of products in each category are[6]:

- Entity Extraction = 14 open source, 50 commercial
- Relationship Extraction = 2 open source, 24 commercial
- Semantic Integration = 6 open source, 15 commercial
- Entity Disambiguation = 2 open source, 8 commercial
- Knowledge Base = 17 open source, 17 commercial
- Visualization = 18 open source, 36 commercial
- Query = 18 open source, 36 commercial
- Analysis = 4 open source, 40 commercial
- Ontology/Data Model = 26 open source, 15 commercial
- Reference Data = 7 open source, 1 commercial

(U) The following trends and directions, by category of tool, can be inferred from the data in Appendix C.

(U) Entity extraction (64 Tools)

(U) Relevant Performance Characteristics:

- Maturity
- Integration with other tools
- Precision
- Recall
- Is it interactive?
- Nature of output (e.g., list of entities or "tagged" text)
- Multi-lingual?
- Ontology-based extraction?

(U) Key findings

- There are lots of vendors in this tool category, but implementation maturity is only "medium" for most.
- Many of the tools (21) also perform Relationship Extraction.
- There are a fair amount of open source tools (14).
- Aerotext, Inxight, and NetOwl are the best known "pure play" entity extractors.
- Many of the entity extraction algorithms have been around for some time; there are some new algorithms but no significant breakthroughs.
- This is a secondary service for most of these tools.

[6] (U) They do not add up, since many products have functionality in more than one category

- Some vendors report achieving precision and recall scores over 90%, but comparative measures of performance are often lacking. All claims of performance should be taken with a grain of salt.
- Customization of the tools for specific domains is difficult, time-consuming, and with unknown resulting performance.

(U) Relationship extraction (26 Tools)

(U) Relevant Performance Characteristics:

- Maturity
- Integration with other tools
- Precision
- Recall
- Is it interactive?
- Nature of output (e.g., list of entities or "tagged" text)
- What types of relationships? (person-to-person? person-to-place and -time?)
- Measure of intensity of relationship?
- Measure of frequency of relationship over time?
- Multi-lingual?
- Ontology-based extraction?

(U) Key findings

- There are a fair number of vendors, but implementation maturity is only "medium" to "early."
- Nearly all (21) also do Entity Extraction.
- There are very few open source tools (2).
- Aerotext, Inxight, and NetOwl are the best known.
- Many of the relationship extraction algorithms have been around for some time; there are some new algorithms but no significant breakthroughs.
- This is a secondary service for many of these tools.
- Performance is typically poor, with precision and recall often as low as 60%.
- Customization of the tools for specific domains is difficult, time-consuming, and with unknown resulting performance.

(U) Semantic integration (21 Tools)

(U) Relevant Performance Characteristics:

- Maturity
- Integration with other tools
- Scalability
- Batch or streaming input?
- Interactive?
- Source of integration rules
- Does the tool overwrite input data or implement the integration rules at time of query/analysis?
- Preserving/retaining provenance
- Relationship to RDF store (Read from RDF? Write to RDF?)

- Integrate both entities (nouns) and relationships (verbs)?

(U) Key findings

- There are many tools that do integration (mostly from the database world), but few are focused on semantic integration.
- Nearly one-third of the tools (6) are open source.
- The tools exhibit a variety of maturity, but few are mature.
- There does not seem to be recognition by commercial vendors of the need for this function.

(U) Entity disambiguation (8 Tools)

(U) Relevant Performance Characteristics:

- Maturity
- Integration with other tools
- Scalability
- Precision
- Recall
- Interactive?
- Multi-lingual?
- Soundex/Phonetic matching?
- Stemming?
- Handles misspelling, data entry errors?
- Requires RDBMS in some cases
- Relationship to RDF store (Read from RDF? Write to RDF?)
- Type of analysis (e.g., imputation, matching algorithms, etc.)
- Industry applicability

(U) Key findings

- Relatively few products perform this function (5 companies with 8 products), and nearly all of them offer other relevant services within the tool as well (e.g., query, analysis, knowledge base, visualization).
- Open source tools exist and tend to be more specialized.
- Some tools are mature and proven, but often only in a non-government context (e.g., Healthcare).
- There is a lack of comparative measures of performance (precision and recall), so claims of performance are often missing or impossible to interpret.

(U) Knowledge Base (34 Tools)

(U) Relevant Performance Characteristics:

- Maturity
- Integration with other tools
- Scalability
- Underlying data model is RDF?

(U) Key findings

- There are quite a few tools available, and half of them (17) are open source.

- All tools also support querying of the knowledge base.
- Most specialize in storage and retrieval of semantic data, and most are based on Semantic Web standards, such as RDF and OWL. However, it is often difficult to discern how much of the standard is supported.
- The performance of the tools, such as load time and query time for some standard queries (like the Lehigh University benchmarks) is rarely given, and at scale this will be an important discriminator.
- While some triple stores can now load up to 40,000 triples per second, the average seems to be around 10,000 per second for up to a billion triples stored[7].

(U) Visualization (54 Tools)

(U) Relevant Performance Characteristics:

- Maturity
- Integration with other tools
- Scalability
- Geographic component?
- Temporal component?
- Ability to link to RDF?
- Ability to filter?
- Ability to drill-down?
- Tied to specific type of analysis?

(U) Key findings

- There are many tools available, and one third of them (18) are open source.
- Most of the open source tools specialize solely in visualization.
- About half of all the visualization tools (27) also do some kind of Analysis (and visualize the results of the analysis).
- A few (11) also perform the Query function.
- Some are fairly mature in terms of implementation.
- Some are tied to visualizing a specific type of analysis (e.g., link analysis).

(U) Query (54 Tools)

(U) Relevant Performance Characteristics:

- Maturity
- Integration with other tools
- Ability to query RDF?
- Support concept search
- Support stemming
- Support soundex/phonetic search
- Language capability

[7] Source: "Measurable Targets for Scalable Reasoning" by Atanas Kiryakov, Ontotext Lab, Sirma Group Corp., 27 November 2007 http://www.ontotext.com/publications/ScalableReasoningTargets_nov07ak.pdf.

- Handles misspelling, data entry errors?

(U) Key findings

- There are many tools available, and one third of them (18) are open source.
- Very few tools (7) specialize solely in Query, which is expected, since they are connected to some storage mechanism, which will define their query approach.

(U) Analysis (44 Tools)

(U) Relevant Performance Characteristics:

- Maturity
- Integration with other tools
- Scalability and computational speed
- Ability to operate on an RDF store
- Requires structured input?
- Nature of output?
- Faceted search?
- Measures of association (e.g., centrality)?
- Imputation of nodes and/or links?

(U) Key findings

- There are many tools, but only a few (4) are open source.
- Most of the tools (29) also do Visualization of the analysis results.
- Some of the tools (17) also do Query.
- Few complex analyses are done.
- There is a mixed maturity of implementation, with some well-established, mature products, but many immature products.

(U) Ontology/data model (41 Tools)

(U) Relevant Performance Characteristics:

- Maturity
- Integration with other tools
- Includes baseline ontology?

(U) Key findings

- There are many tools of relative maturity, since the community has focused on data modeling to date.
- The majority of tools (26) are open source.
- Nearly all are specialized tools for building and managing ontologies.
- There is little information on interoperability, but experience has shown that most tools do not adhere completely to the standards.

(U) Reference data (8 Products)

(U) Relevant Performance Characteristics:

- Comprehensiveness
- Update frequency

- Support multiple languages
- Geographic translation (place name to lon-lat)
- Export format

(U) Key findings

- Most (7) are government databases that are openly available at no cost.
- There are probably many more reference data sets than we present herein.

5. (U//FOUO) Government Systems for Catalyst (U)

(U//FOUO) The following government systems were identified and described under the study:

- **Aether** from Office of Naval Intelligence (ONI)
- **APSTARS** from National Security Agency (NSA)
- **BlackBook2** from Intelligence Advanced Research Projects Activity (IARPA)
- **Common Ontological Data Environment (CODE)** from Joint Warrior Analysis Center (JWAC)
- **Future Text Architecture** from JWAC
- **Harmony** from National Ground Intelligence Center (NGIC)
- **Information Extraction/Structured Data Analysis (IE/SDA)** from Central Intelligence Agency (CIA)
- **Intelligence Integration Cell (IIC)** from National Counter Terrorism Center (NCTC)
- **K-Web (GeoTaser + Knowledge Miner)** from National Geospatial-Intelligence Agency (NGA)
- **Large Scale Internet Exploitation (LSIE)** from DNI Open Source Center (OSC)
- **Metadata Extraction and Tagging Service (METS)** from Defense Intelligence Agency (DIA)
- **Pathfinder** from NGIC
- **Quantum Leap** from CIA
- **Savant** from National Air and Space Intelligence Center (NASIC)
- **Vocabularies for the IC to Organize and Retrieve Everything (VICTORE)** from CIA

(U//FOUO) It is interesting to note that nearly every major IC Agency has recognized the need for Catalyst functionality (although they do not call it as such) and have allocated resources to developing capabilities to implement one or more of the functions. These systems are described in summary form in Appendix D. Note that one of the aspects that is described in Appendix D is the COTS, open source, or GOTS products that each one of these systems use, which are in turn described in Section 4 and Appendix C.

(U//FOUO) There are several systems at NSA, in addition to APSTARS, that are relevant to Catalyst. These were examined and described, but they are not included herein due to security considerations. A classified appendix to this report contains these descriptions.

(U//FOUO) The descriptions and associated interviews with many of the managers and developers of the systems resulted in the following observations.

- (U//FOUO) Entity extraction, and to a lesser degree relationship extraction, is a well-funded, active area of development across the IC. There are significant programs at CIA, DIA, NSA, and some of the other agencies. The persistent issues with these programs seem to be (1) quality of output, (2) throughput, and

(3) difficulty of development. The first issue, quality of output, deals with the accuracy (both precision and recall) of the entity extraction, and the difficulty of even assessing how accurate the extraction is. Most programs have stated that entity extraction is generally higher quality than relationship extraction, and in fact some programs are not even doing more than entity extraction. But in most cases little hard data is available about the quality of extraction, due to the difficulty of defining metrics and measuring performance. The second issue is how many documents can be run through the extraction process per unit time. It was stated to be a limiting factor in several implementations. The third issue is how expensive and difficult it is to stand up services to do extraction. The underlying COTS are often expensive, especially if licensing is organization-wide, and the number of hours of development and tuning is significant. Many programs have determined that to get high quality extraction the products must be tuned to the particular types of documents and domain of analysis; this is often a long and complex process.

- (U//FOUO) Once information is extracted from documents or other data, different agencies take different approaches as to what to do with it. Some, like METS, are a stateless service that extracts information and provides the extraction back to the requester and does not persist the data. (Note that earlier METS did persist the data; it is now part of the DoDIIS Data Layer, not METS, but the capability is still extant.) Most other extracted information is stored according to a data model that captures the salient information, in various forms corresponding to differing levels of formal semantics. One good example is the Common Representation Format of the CIA's IE/SDA program. Very few organizations have gone as far as an RDF triple store with OWL support. More often than not, the results are stored in a relational database management system, since this is robust, scalable, well understood technology. Few systems have scaled up or used the more complex semantic factors of the data to need any other storage approach, nor has the use of inference been widespread, part of the justification for the semantic model. Basically, a compelling need for any persistent model beyond the relational model has not yet been demonstrated in an operational setting.
- (U//FOUO) Semantic harmonization and integration has been attacked in many of the systems, but mostly in an ad-hoc way. Often custom code has been written that maps data from various sources into the data model of the system. There are few cases of doing formal mappings from schemas or ontologies into a common schema or ontology, and then translating the instance data according to that mapping. Of course, when custom code is written to translate it takes into account the schema or ontology of the data source, but not in a formal process. The differences in approaches have implications for maintenance. Surprisingly little has been done in integration. Mostly additional property values are kept, along with the source of the information. This might be indicative of the need for keeping all original data to use for analysis.
- (U//FOUO) Disambiguation is understood to be important to most of the systems that integrate data in some fashion. Some systems have written custom code to do the disambiguation function, especially when based on names. Others have used

commercial products, but few systems have had success with this approach. A program such as Quantum Leap, which is more mature in disambiguation than most, has gone through several approaches before settling on the one currently used. One very important issue in integration and disambiguation is the security aspect, especially when US persons are involved. Most programs have not yet dealt with this aspect, except for Quantum Leap. (This is not to say that programs have not had to deal with the usual security requirements of any system that operates on JWICS or other Top Secret networks.) Quantum Leap has developed some unique procedures to ensure adherence to requirements of dealing with US persons.

- (U//FOUO) No government system has yet dealt with integration of entity knowledge bases. Rather, they have integrated relational databases or other sources of information along with extracted information. This situation is due to the paucity of entity knowledge bases with which to integrate.

6. (U//FOUO) Conclusions and Recommendations (U)

(U/FOUO) The area that Catalyst is envisioned to address is an important one for the Intelligence Community (IC). Catalyst is trying to move beyond the current IC push, where to "share" means to make all resources (documents, images, web pages, etc.) available on the same networks accessible by all analysts. Whereas this is a necessary condition for sharing, it often results in an analyst in one organization within the IC who used to be drowned by all the available resources in his or her organization on a topic now being drowned by all the available resources in *all* organizations across the IC on the topic. This will not make his or her job any easier. Fundamentally, by sharing all resources we have increased *recall*, but even if *precision* were high, the sheer number of results from many queries is too much for an analyst to absorb.

(U//FOUO) Many IC organizations have recognized this problem and have programs to extract information from the resources, store it in an appropriate form, integrate the information on each person, organization, place, event, etc. in one data structure, and provide query and analysis tools that run over this data. Whereas this is a significant step forward for an organization, no organization is looking at integration across the entire IC. The DNI has the charter to integrate information from all organizations across the IC; this is what Catalyst is designed to do with entity data. The promise of Catalyst is to provide, within the security constraints on the data, access to "all that is known" within the IC on a person, organization, place, event, or other entity. Not what the CIA knows, then what DIA knows, and then what NSA knows, etc., and put the burden on the analyst to pull it all together, but have Catalyst pull it all together so that analysts can see what CIA, DIA, NSA, etc. all know at once. The value to the intelligence mission, should Catalyst succeed, is nothing less than a significant improvement in the analysis capability of the entire IC, to the benefit of the national security of the US.

(U) Recognizing the importance of Catalyst-like capabilities, significant work has gone on to develop capabilities in the commercial and open source product worlds, in the government systems world, and in academia. This report has addressed what the commercial/open source and government worlds have to offer to Catalyst, so that the development of Catalyst can take full advantage of the work in this area. Over 230 commercial and open source products were discovered and described, and a sense of the direction of the commercial world in the functions relevant to Catalyst were inferred. Approximately 20 programs in the IC were identified and described, and observations from these programs were collated and described.

(U) What emerged from this collection and analysis is a nascent technical area that has great promise, but is still in its infancy. Whereas pockets of commercial/open source or government systems are mature, many more are in development without a proven track record. The next few years are bound to be very interesting to see what transpires, and to see if the promises of the area are met.

(U) One recommendation of this study is to continue to perform the kinds of tracking and evaluations that have been done for this report. New companies, new products, enhancements to products, companies ceasing operation or pulling a product off the market, and other developments should be tracked to continue to provide the reference

data (such as in Appendix C) and examined to see if the observations of Section 4 remain in effect. Likewise, government programs develop new capabilities, scale up to greater processing capacity, incorporate new functionality, operate on new resources and in new analysis domains, and assess their own performance relative to analysis capabilities as time goes on. These new developments should be tracked to provide the reference data (such as in Appendix D) and examined to see if the observations of Section 5 remain in effect.

(U//FOUO) Another recommendation from this study is to empower appropriate groups within the IC to standardize languages and ontologies. That is, no harmonization would be necessary if all Entities Knowledge Bases used the same language to express their entities with attributes and relationships, and if the semantics of the entities, as represented in the ontology of the Knowledge Base, were coordinated. Note that we did not say that the ontologies should be the same, since the Entities Knowledge Base will need to serve a local need, and different organizations have different local needs. However, where the ontologies of the individual Entities Knowledge Bases need to represent common entity classes, like persons, organizations, events, etc., and common attributes and properties of these classes, standards should be developed so that harmonization processing is minimized or entirely eliminated. It is not unexpected that the developments of Entity Knowledge Bases around the IC have been uncoordinated to date, since in many cases the developers were not aware of other efforts, and in any case there was not, and indeed still is not, a common core ontology for common classes. Processes should be stood up to develop and manage core ontologies, and all local ontologies should be rooted in the core ontologies.

(U//FOUO) In this study we have purposely not addressed the security requirements of an eventual Catalyst system, as it is outside the scope of the study given the resources available. However, any eventual Catalyst implementation will have to deal with some serious security concerns. The pedigree of resources was touched upon herein only, but only superficially, and how the pedigree and security work together is a big issue. Thus another recommendation is to elucidate the security requirements that are unique to a Catalyst implementation and analyze how these requirements impact the functionality and design of the system. One known issue in this area is how much entity data in the Integrated Entities Knowledge Bases needs to be protected, and how this impacts what information can be queried (and by whom) and what information can be provided back to the original Entities Knowledge Bases.

(U//FOUO) A last recommendation is with regard to the implementation of Catalyst-like capabilities, that has only been touched on briefly in this report. For interoperability and leveraging capabilities, a software architecture for Catalyst-like capabilities across the IC should be developed and services of common concern stood up where possible, in a Service Oriented Architecture. There is clearly duplicative effort going on within the IC today, and while this is healthy at this point, since many issues remain about technical approach, performance, etc., once some of these issues are settled sufficiently the opportunity exists to leverage capabilities of one organization's implementations for other members of the IC. Such sharing of services deserves consideration in an eventual Catalyst system.

(U) There is no doubt in the minds of the authors that there is significant information that was missed, simply due to resource constraints in the study. One way to increase the coverage of the reference data on which the observations are made is to distribute the task of data collection to the organizations that comprise the IC, and even the commercial companies that supply or are interested in supplying products to the IC. The more that many people can collaborate on providing data, the more likely it is that coverage improves. Thus, one recommendation is to find mechanisms for discovery of other commercial and open source products and government programs, and let them self-describe in some form so that the collection resources needed are minimal. Then whatever resources are available can be concentrated on the analysis of the data, thus improving the observations.

(U) Appendix A. Terminology[8] (U)

(U) **Entity**: a representation of a thing in the real world, either concrete or abstract[9]. Each entity is an instance of a class, where these classes form a hierarchy. The class hierarchy is a formal "is-a" or specialization/generalization hierarchy. Entities can be a member of more than one class. Entities have *properties*, whose values can be data or whose values can be other entities. The former properties are called datatype properties, or sometimes *attributes*, while the latter properties are called object properties, or sometimes *relationships*. Entities are sometimes called *knowledge objects*, *semantic knowledge objects*, *semantic objects*, or simply *objects*.

> Examples[10] of classes are `person`, `Organization`, `Place`, `Event`, or `Artifact` (a man-made physical thing). The `Person` class, for example, might be a subclass of `Agent` and a superclass of `MilitaryPeople`, `Citizens`, `Resident Aliens`, `Politicians`, `ComputerScientists`, etc. Examples of entities are `John Doe` as an instance of the class `Person`, the `SalvationArmy` as an instance of the class `Organization`, or `ArlingtonVirginia` as an instance of the class `Place`. Examples of datatype properties are the `Name` (a string) or `Age` (a positive integer) of a `Person`, the `Name` (a string) or `NumberOfMembers` (a positive integer) of a `FraternalOrganization`, the `Population` (a positive integer) of a `GeopoliticalEntity`, the `DateOfOccurrence` (a date) of an `Event`, or the `Length` (a floating point number) of a `Car`. Examples of object properties are the `Father` (a `Person`) or the `Employer` (an `Organization`) of a `Person`, the `Affiliation` (an `Organization`) of an `Organization`, the `mayor` (a `Person`) or `StateContainedIn` (a `GeopoliticalEntity`) of a `City`, the `Participants` (a group of `Persons`) of a `MeetingEvent`, or the `Owner` (a `Person`) or `Location` (a `GeopoliticalEntity`) of a `House`. Note that in each case a property has three parts: an entity (of a specified class), the name of the property, and the value of the property, which is either a string, number, etc. (a datatype) or another entity (of a specified class).

(U) **Entity extraction**: the identification and classification of entities embedded in some kind of unstructured data, such as free text, an image, a video, etc. "Identification" means delimiting the entity in the data (although sometimes this is not possible), and "classification" means assigning a class in the class hierarchy to the entity. Sometimes seen as part of entity extraction is the identification of datatype properties of the entity.

---

[8] (U) A word on representation: many representations could be used for a system such as envisioned herein, such as a relational model (as implemented in an RDBMS) or a spatial model. However, only a semantic graph has demonstrated the potential to scale and represent the information under consideration. Therefore, this memo assumes the enterprise-level representation is a semantic graph.

[9] (U) We use the term "real world entity" to refer to the actual thing in the real world.

[10] (U) We adopt the usual semantic web practice of naming classes, properties, and instances using Camel Case (see http://en.wikipedia.org/wiki/CamelCase).

> For example, some free text may include "... Joe Smith is a 6'11" basketball player who plays for the Los Angeles Lakers..." from which the string "Joe Smith " may be delineated as an entity of class `Athlete` (a subclass of `People`) having property `Name` with value `JoeSmith` and `Height` with value `6'11"` (more on this example below). Note that it is important to distinguish between an entity and the name of the entity, for an entity can have multiple names (`JoeSmith`, `JosephSmith`, `JosephQSmith`, etc.).

(U) **Entity disambiguation**: the association of two entities extracted from data as being two instances of the same real-world entity. The resolution can be between two entities extracted from the same resource (such as a single document) or it can be between an entity extracted from a resource and an entity from another resource (such as two documents) or it can be between an entity extracted from a resource (such as a single document) and entities saved in a knowledge base (see below). This process is also called "co-reference resolution" and "identity resolution."

(U) **Relationship discovery**: the identification and classification of object properties (*relationships*) embedded in some kind of unstructured data, such as free text, an image, a video, etc. "Identification" means delimiting the entity in the data (although usually this is not possible, so is rarely done), and "classification" means assigning a specific property to the relationship (that is, not simply saying that two entities are related, but saying *how* they are related). Since relationships are always between two or more entities, relationship discovery has to be done in concert with entity extraction (although it is possible for a relationship to be between unknown entities), whereas entity extraction can be done without relationship discovery. To be consistent with the term "entity extraction" and to reflect how relationships are derived from resources just as entities are, this process should more accurately be called "relationship extraction," but this is not a common term.

> To continue the example above, the entity with `Name JoeSmith` has property `MemberOf` having value an entity of class `SportsFranchise` (a subclass of `Organization`) with `Name Lakers`, which, in turn, has property `LocatedIn` having value a `City` with name `LosAngeles`. Note that it is not obvious where to delineate this property, which is why relationships are normally associated with the data and not delineated in the data.

(U) **Knowledge base**: a collection of entities (instances). Each entity is described in terms of the class of which it is a member, and the property values that are known about the entity (that is, the values that have been extracted). Since much of the information stored is in the form (entity, property, value), these are called *triples* and the knowledge base a *triple store*[11]. One especially useful way to describe such a collection of entities

[11] (U) Actually, in most knowledge bases the triples also contain metadata, such as the resource (the document or video or ...) from which the values are extracted or who validated the information or the classification of the data. As such, a more accurate term for these knowledge bases are *quad stores*, where each datum is a triple of an entity's property with value and the associated metadata.

and their properties is as a *semantic graph*, with each entity (instance) a node of the graph and the edges of the graph being named properties connecting the nodes.

> To continue the example, one entry in the knowledge base is the entity of class `Athlete` with (datatype property) `Name` having value `JoeSmith`, another is the entity of class `SportsFranchise` with `Name` having value `Lakers`, and another is an entity of class `City` having value `LosAngeles`. If each of these is viewed as a node in a graph, then an edge connecting the node (entity) with `Name JoeSmith` to the node with `Name Lakers` is named `MemberOf` and the edge connecting the node with `Name Lakers` to the node with `Name LosAngeles` is named `LocatedIn`. Such edges, corresponding to relationships (object properties) and have a direction; for example, `JoeSmith` is a `MemberOf` the `Lakers`, but the `Lakers` are not a `MemberOf JoeSmith` (there may be an inverse relationship, such as `HasMember`, that is between the `Lakers` and `JoeSmith`.). Thus, the entire knowledge base is a *directed semantic graph*.

(U) **Ontology**: the definitions of the classes and the properties of the classes is called an *ontology*. Properties are inherited, so that a class B that is a subclass of class A has all the properties of class A plus others that are unique to class B. An ontology also includes statements about classes and properties, such as that one property is the inverse of another property. Often the ontology is also stored in the knowledge base[12].

> As an example of inheritance, say that there is one class called `Vehicles`, a subclass of `Vehicles` called `WheeledVehicles`, and a subclass of `WheeledVehicles` called `Automobiles`. A property of the class `Vehicles` may be `MaximumSpeed`, since this property applies to all vehicles. A property of `WheeledVehicles` may be `NumberOfWheels`, which is appropriate for this class but not some other subclass of vehicle (such as `TrackedVehicles`), and this class also inherits the `MaximumSpeed` property from its parent class. A property of the class `Automobiles` may be `NumberOfDoors`, which is appropriate for this class but not some other subclass of `WheeledVehicles` (such as `Motorcycles`), and this class also inherits the `MaximumSpeed` and `NumberOfWheels` properties from its parent class.
>
> As an example of statements about properties, say we have an ontology of people. One object property of a `Person` may be `ParentOf`, and another `ChildOf`, and a third `FriendOf` (all three properties of the class `Person` have as value another instance of the class `Person`). We can state that `ParentOf` is the inverse of `ChildOf`, and then if we know that `John` is the `ChildOf Bill`, we do not have to explicitly state that `Bill` is the `ParentOf John`, since it can be inferred from the fact that the two

[12] (U) Note that some people include in the definition of an ontology some "base" instances, or even all instances. It is most common to use the term “ontology” to only refer to the class hierarchy, the properties, and statements about the instances of the classes and properties, and not to instances.

properties are inverses. Likewise, we can state that `FriendOf` is symmetric, and then if we know that `John` is the `FriendOf Harry`, we do not have to explicitly state that `Harry` is the `FriendOf John`, since it can be inferred from the fact that the two properties are symmetric.

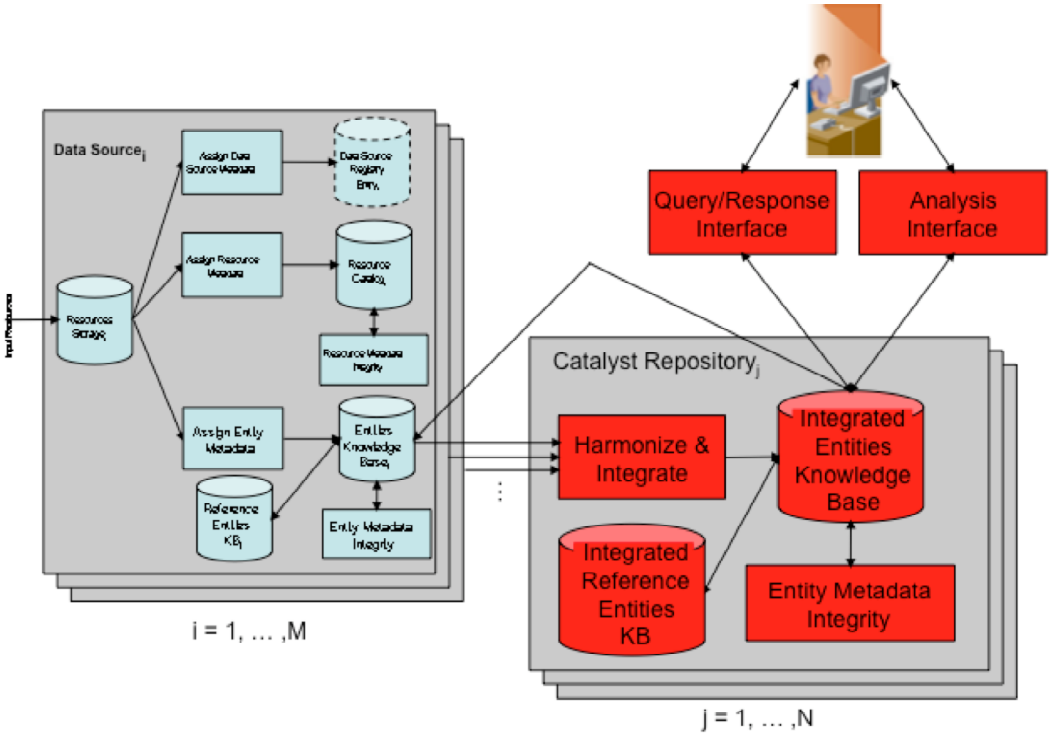
(U) **Pattern**: a (partially) uninstantiated set of two or more entities, with specified relationships among them (including the "unknown" relationship). The simplest pattern is two entities with one relationship between them, where at least one of the entities and/or the relationship is uninstantiated. Patterns can become arbitrarily large and complex. Some people would include in the definition of a pattern conditionals, branches, recursion, etc.; there is not a well-accepted definition of pattern to know whether or not to include these constructs.

A simple pattern could be: `Person Owns Automobile`, where both `Person` and `Automobile` are uninstantiated. It can be instantiated by any specific instance of `Person` who owns a specific instance of `Automobile`, for instance `JoeSmith Owns` an instance of the class `Automobile` with the `Manufacturer` property having value `Lexus` and the `LicensePlate` property having value `VA-123456`. Another simple pattern could be: `Joe Smith Owns Automobile`, or `Person Owns` an instance of the class `Automobile` with `Manufacturer Lexus` and `LicensePlate VA-123456` or even `JoeSmith` has-unknown-relationship-with an instance of the class `Automobile` with `Manufacturer Lexus` and `LicensePlate VA-123456`. In these last three examples, one of the entities or the relationship is uninstantiated. Note that `JoeSmith Owns` an instance of the class `Automobile` with `Manufacturer Lexus` and `LicensePlate VA-123456` is **not** a pattern, for it has no uninstantiated entities or relationships. A more complex pattern could be: `Person Owns Automobile ParticipatedIn Crime HasUnknownRelationshipWith Organization HasAffiliationWith TerroristOrganization`. Any one or more of the entities and the has-unknown-relationship-with relationship (but not all) can be instantiated and it would still be a pattern, such as `JoeSmith Owns Automobile ParticipatedIn Crime PerpetratedBy Organization HasAffiliationWith HAMAS`. An example of recursion in a pattern is: `Person Owns Automobile ParticipatedIn Crime HasUnknownRelationshipWith Organization HasAffiliationWith (HasAffiliationWith (... TerroristOrganization)))`, where the depth of `HasAffiliationWith` may be specified (no more than 4 deep, for example). Instantiation of patterns can be by any instance of the class specified or by an instance of one of its subclasses, so that if the subclasses of `Automobile` are `ForeignMadeAutomobile` and `AmericanMadeAutomobile`, and an instance of the class `Automobile` with `Manufacturer Lexus` and `LicensePlate VA-123456` is an instance of `ForeignMadeAutomobile`, it still is an instantiation of the pattern.

(U//FOUO) Appendix B. Detailed Description of Functionality (U)

(U) We assume that this generic processing starts with unstructured and semi-structured data, such as documents, images, videos, audios, signals, measurements, etc., as well as structured data, that are collected from a wide variety of sources by a variety of methods. We use the term *resource* to include all of these input data types. We use the term *resource* to include all of these input data types[13]. The objective of the advanced intelligence processing is to identify the ***entities***—people, places, organizations, events, etc.—in the resources and what the resource says about the entities (the attributes of entities and the relationships among them), and make this information available to users (generally, intelligence analysts) so they can retrieve information and detect patterns of interest to their analysis mission.

(U) At a high level, there are three steps to the kind of intelligence processing related to Catalyst[14]: (1) *describing resources and making them capable of being processed,* (2) *semantically integrating entities of interest to a specific task* (including disambiguation of these entities), and (3) *processing the entities to produce some conclusion of interest*. The Figure below summarizes the steps; each step is expanded upon below.



[13] (U) In the remainder of this report we assume that the resources are documents, although many of the issues and approaches also apply to other kinds of resources.

[14] (U) We are only describing analysis, not collection, although there clearly should be a connection between the two that exists today only in rudimentary form. An analyst should be able to express his or her entities of interest, and not only should the currently held resources be searched, but if there is not sufficient information (which is difficult to determine automatically) a collection request for more resources should be initiated.

(U) The first step is done mostly automated, although some of the description of resources may be manual. The second step is usually initiated by an analyst/user, who describes in some way what entities are of interest to him or her at the moment to support his or her analysis tasking, and then the finding of entities of interest and integrating them is automated. The third step is mainly manual today, although there are tools that significantly support the processing and production of conclusions. It is the belief of many (including us) that more automation must be done in the final step, since the volumes of data preclude manual processing. An overlay on all of these steps is that analysis is usually not done by a single individual, but by many individuals, so that collaboration in the three steps is important. Also, it should be noted that the result of analysis is often additional resources that are described and made available for processing, so there are feedback loops in the processing steps.

(U) Describing Resources for Processing

(U) In order to process the entities in resources, they need to be explicit in a structured form. Some resources are naturally structured (data that is typically in a relational database management system, for example), and so are already in a form that capable of being processed. Other resources are either unstructured or semi-structured[15], such as a document, an image, a video, or a signal. In order to make resources capable of being processed by a Catalyst system, structure in the form of *metadata* must to be added to them. It is a combination of the original resource and this metadata that is persisted, and the metadata indexed for search[16].

(U) In general, metadata that describes intelligence resources falls into three categories: *descriptive metadata*, *structural metadata*, and *content metadata*[17]. *Descriptive metadata* provides information about the resource as a whole, such as title, authoring agency, security classification, or date. For documents, the most common *descriptive metadata* approach is that of the Dublin Core[18] from the library science community. (There are approaches, both within and without the government, for metadata for other types of resources, such as NITF or ISO/IEC 15444-1 for imagery.) *Structural metadata* describes how a resource is laid out for rendering. This type of metadata has little to do with processing entities, so is not discussed further herein, although it is certainly important for an overall intelligence processing system. *Content metadata* describes what the resource "is about;" it can relate to the resource as a whole, such as the topic or

[15] (U) Unstructured data, such as documents or images, have no inherent structure that describes them, while semi-structured data has an unstructured part–the text of the document or the image–and a structured part that describes the unstructured part, such as the author, title, date of publication, etc.

[16] (U) Depending on the processing, it may operate on the original resource in addition to the metadata, such as text keyword searching.

[17] (U) These terms are not widely agreed upon.

[18] (U) See http://dublincore.org/.

geographic area that the resource is about[19], or it can relate to the details inside the resource, such as the specific entities mentioned in the resource and what the resource says about the entities (that is, attributes of or relationships among the entities). This latter content metadata, the entities and their relationships, is sometimes referred to as "deep content." In this study we are only concerned with content metadata.

(U) With reference to the Figure above, we define a *Data Source* as a collection of resources plus any metadata about the resources, and tools to process the resources, such as search, analysis, etc. Each Data Source receives new resources by some mechanism. The resources are stored persistently for retrieval, and three metadata processing steps are performed. Each Data Source contains metadata to describe the Data Source as a whole, used for discovery of Data Sources that may be of use to a particular intelligence processing task. This metadata is provided to a Data Source Registry for indexing and search. In addition, all resources are assigned resource metadata, which includes descriptive metadata and content metadata that is about the resource as a whole. This resource metadata is stored in a resource catalog, so it can be searched and relevant resources retrieved. These processing steps are *not* the subject of this study, although they are required (in some form) for use of the resources in the Data Source for intelligence processing.

(U) All resources also are assigned entity metadata; that is, the entities in the resource are identified, delimited, and assigned to a class and, where possible, the attributes and relationships among the entities in the resource are identified. The entity metadata is stored in an Entity Knowledge Base. The Entity Knowledge Base often includes *reference entities*–representations of well-known and accepted real world entities–that are not derived from any resource, but input into the entity knowledge base by some other mechanism.

(U) There are two major issues regarding metadata. The first is *how* it gets assigned[20], and the second is *what* metadata is assigned. It is the fervent hope and the naïve assumption of many people that high quality metadata can be assigned by some magic program running fully automated (right out of the box), and many commercial companies sell their products with this promise in mind. The reality is that there is currently no way to assign high quality metadata automatically to a broad set of resources; either the quality is mediocre to poor, or some manual process must also be included, and even then the quality is often not very good, or the domain over which the metadata is assigned is severely limited. We don't expect this situation to change in the near future. Many approaches have been taken to assigning content metadata, such as clustering techniques and other statistical methods that use co-occurrence of words in a document to determine the overall topic, or entity and relationship extraction approaches to derive the deep content of a document. None of these approaches has been shown to provide high quality metadata, although few serious benchmarks with ground truth have been done to

[19] (U) Dublin core includes such content metadata.

[20] (U) We am using the term "assigned" to denote that it may be automated or manual, but in either case the end result is that there is metadata associated with the resource. Note that we also are not addressing herein whether the metadata so assigned is made a part of the original resource or is separate (e.g., in a "metacard"), for these are implementation issues and not functionality issues.

generally validate this assertion[21]. If, instead, a manual metadata assignment process is taken as the approach, the tools developed to support the user in making the assignments have generally been difficult and time-consuming to use, and most people have shied away from using them, or have fought the directive to use them. This comment applies primarily to deep content metadata; it has been somewhat more successful to develop and use tools for manual assignment of descriptive metadata and resource-level content metadata, including security attributes. But even in this area significant improvements could be made.

(U) The previous discussion had to do with the process of assigning metadata, but a significant additional issue is *what* metadata to assign. Some organizations take the minimal approach and only assign a small number of key elements, while others assign many more. Often the meaning of these elements is not clear between organizations. If, for example, one organization expresses several dates in its metadata (production date, publication date, cut date, etc.) but another organization only expresses one date, which is it? How do we use resources from both organizations together? How do we even interpret the one date from the second Data Source if that is the only resource that we are interested in? In addition to these issues, within the IC there are few, if any, common controlled vocabularies. For example, for the *subject* or *topic* of a resource, there have been many different local (within an organization or a part of an organization) controlled vocabularies, such as DIA's IFCs, OSC's, or ICES's topic directory, and many organizations today use the NIPF, the National Intelligence Priority Framework, as the controlled vocabulary for subject. There are several problems with these approaches. Foremost among them is that if an organization develops its own version of a subject controlled vocabulary, which is appropriate for serving its customer's needs, it is often not clear how this vocabulary should be interpreted by others outside the local customer base. If, as is usually the case today, the meaning of the vocabulary is implicit, or explicit but not formalized, then manual intervention will be needed to interpret the metadata, and it is likely that there will be lingering interpretation issues among vocabularies that will limit the ability to use the data across organizations. (Also, NIPF is not an appropriate subject vocabulary since, as a priority framework, it changes as the national security situation changes, and the subjects of resources do not change. The reason it is being used, in our opinion, is that there is no good alternative that is common across the IC.[22])

(U) The same story holds for entities and relationships among them. If one tool determines that the entities in a resource are, say, a *person*, *place*, or *thing* (common for out-of-the box COTS entity extractors), while another tool determines if a person is a particular type of person but doesn't know about places, or another tool determines only geospatial entities, then each may serve its own local use, but there will be the same interpretation issues as when trying to use more than one subject vocabulary. And even if both tools find, for example, places, but one determines geopolitical and geophysical

[21] (U) A fact exploited by the sales and marketing departments of most commercial vendors of such products.

[22] (U) Thanks to Dave Roberts of CIA/Data Architecture for helping us understand and appreciate this issue.

features while the other only determines geopolitical features, then how do we use both metadata elements together?

(U) The real issue here is how people or tools assign content metadata in a way that it is widely usable. It is tempting to say that the way the IC is going to solve this problem is to standardize on one content metadata element set with one common controlled vocabulary. Not only is this not a good solution if it could be implemented, since each Data Source needs to address its local customer base that may need particular metadata, but there are many social and organizational reasons why this will not succeed. Indeed, companies and government organizations have tried this approach in the past[23], with little success.

(U) An alternate approach that is more likely to succeed is for the metadata element sets and their associated vocabularies and meanings to be *explicit* and *formalized*. Then it is possible for the metadata to be interpreted unambiguously (or at least, with higher fidelity than if they were not explicit and formal), and, most importantly, by computers, not people. This last point is worth expanding. If the meaning of the metadata elements and their vocabularies are implicit (i.e. in the head of the developers of the Data Source), or explicit but informal (such as in a data dictionary, which is written in a natural language, such as English, and thus not computer understandable[24]), humans may be able to interpret their meaning with a fair degree of accuracy given their intelligence and world knowledge[25], but sharing of the metadata widely to perform the kinds of intelligence processing needed in today's world requires processing relevant metadata by computers, not people, due to the enormous volumes. The only way that a computer program can find relevant entities and utilize them for intelligence analysis is for the resource's content metadata to be understandable to that program, and this means that the meaning of the metadata must be explicitly and formally stated.

(U) The current thinking related to the means by which this metadata is made understandable to computers is that each metadata element set be described as a component of an ontology[26], and this ontology be available as a url on the same network that the resources are on. Then there are means by which content metadata can be understood and integrated by computers without human intervention, or at least with only human intervention at the ontology level (rather than at the specific entity level). Then, for each Data Source accessed, its ontology can be understood and mapped to some

---

[23] (U) Mainly with common database schemas.

[24] (U) We really mean the data dictionary documentation, rather than the DBMS data dictionary. If we include the DBMS data dictionary, then it is explicit and formal, but there are issues about expressibility of the language used to capture the dictionary.

[25] (U) Or they may not. Just because a human is doing the interpretation does not ensure consistency. There is no doubt that humans can do deeper reasoning than computers, but natural language is inherently ambiguous, and unless data dictionaries and other descriptions of metadata elements are complete and adhered to, the potential for misinterpretation will remain.

[26] (U) In this context the term "ontology" is construed to be in the sense of Deborah McGuiness in "Ontologies Come of Age," MIT Press. In Dieter Fensel, Jim Hendler, Henry Lieberman, and Wolfgang Wahlster, editors. Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential. MIT Press, 2003. This definition admits to, for example, formal taxonomies.

common ontology to support processing across the part of the IC that can use the metadata.

(U) With reference to the Figure above, as each Data Source receives new resources, they are assigned entity metadata; that is, the entities in the resource are identified, delimited, and classified (assigned a class from among the known set defined by the ontology), and, where possible, the relationships among the entities in the resource are identified and classified (assigned properties[27] from among the known set defined by the ontology). The process of assigning entity metadata is often called *entity extraction*, a term used by the commercial world in describing their products. In this report the term is meant to actually encompass entity identification, entity class evaluation, entity attribute assignment, and entity relationship assignment.

(U) This entity metadata is stored in an Entity Knowledge Base. The Entity Knowledge Base often includes reference entities that are not derived from any resource, but are input into the Entity Knowledge Base by some other mechanism. These reference entities can be thought of as well-known and accepted entities representing things in the real world that are already well described. In spite of the way it is shown in the Figure, the Reference Entity Knowledge Base is in reality part of the Entity Knowledge Base, but containing these special entities. The management of reference entities is usually different than other entities, so that, for example, if a resource provides a property value of a reference entity that is in conflict with that of the reference entity, this value is not given the same weight as if it were not a reference entity. Specifically, such a conflicting value might be flagged for consideration by an analyst maintaining the reference entity, but it would not be included in the reference entity property value automatically.

(U) One important issue that the Entity Knowledge Base must support is the referencing of entities. That is, applications (including authoring tools) should be able to reference entities in a persistent, unique, global way, so that there is no ambiguity in the reference (i.e., there is no ambiguity in which entity is meant). The mechanism envisioned to accomplish this requirement is to assign a GUIDE = Globally Unique IDentification for Entities to certain entities. The GUIDE is akin to a BE number for fixed facilities, allowing unambiguous reference to the entity in documents, etc. The GUIDE is just a special case of a URI = Uniform Resource Identifier[28]. It cannot be a URL = Uniform Resource Locator, since it must be able to be referenced outside of the web on which the GUIDE was assigned, such as in a document. Not all entities will get a GUIDE (although all will have a URL, since they will be stored as a node of a semantic graph on a network), since many entities may be too uncertain (as to their connection to an entity in the real world) to justify assigning it a GUIDE. It seems of value to assign GUIDEs to instances of certain classes, such as people and organizations, but not to all entities, and only to those instances that are sufficiently well known and of interest. At this point it is not clear what criteria should be used to determine which classes and instances get GUIDEs and which do not. A GUIDE will be assigned only to certain instances of a

[27] (U) Here the term "relationship" can include both an attribute of an entity, such as a person's date of birth or a city's population, and a property of an entity, such as a person's father or a city's state.

[28] (U) See http://www.w3.org/Addressing/.

class. We use the term *Master Entity* to indicate an entity for which a GUIDE has been assigned; these will be the entities for which it is important to be able to reference. Clearly, the IC will need to implement management processes to determine which entities are declared Master Entities, how GUIDEs are assigned to them, and how the property values of Master Entities are updated[29]. The assignment of a GUIDE may only occur in the integration process described below, not by any individual Data Source.

(U) This processing—assigning Data Source level metadata, assigning resource metadata, and assigning entity metadata—and the persistent storage of these metadata elements, prepares the Data Source for serving both its local needs and for integration across the IC.

(U) Semantically integrating entities

(U) Once the Data Sources have been processed to derive associated Entity Knowledge Bases, Catalyst will provide an integration of these entities for processing across the Data Sources. The objective of this integration is to enable the intelligence analysis on "all we know" about each entity, which implies integrating the entity data from various Data Sources. The means by which such integration is done is to partition the Data Sources, and then integrate the Entity Knowledge Bases from each set of Data Sources into a common Integrated Entity Knowledge Base, part of a Catalyst Repository in the Figure. There will be more than one such Repository; the Data Sources will be partitioned into Repositories as described below.

(U) As shown on the Figure, the data must be harmonized and integrated before it is stored in the common Integrated Entities Knowledge Base. Harmonization is a step that brings entities into a common semantics. It is not sufficient to simply write all the entities from the Data Sources into one repository, because the way they describe their entities may not be commensurate. The entities might not be in common semantics from their Entity Knowledge Bases due to factors such as scaling and unit differences, different levels of granularity, different definition of concepts, etc. It is as if we had a room of Russians, Italians, Croats, Persians, and Argentineans, and we asked them for the information they know on a certain person in English. Not only will all the people not understand the question, but if they did and answered it, we would not understand the answers. What is needed is for the question to be translated into Russian, Italian, Croatian, Farsi, and Spanish, the question in the appropriate language asked of each person, and the answers translated back into English. The equivalent for entities in the formal languages in which they are stored (in their Data Source's Entity Knowledge Base) is that the ontology that describes the meaning (semantics) of the entity classes and properties may be different from the ontology of another Data Source's Entity Knowledge Base. The differing ontologies are akin to the differing spoken languages in the analogy above. A trivial example of harmonization is if the ontologies use different units for some property of a class. For example, if the ontology for Data Source$_1$ has a

[29] (U//FOUO) A particular problem the IC has with respect to GUIDEs is how they get managed across security domains. The term "globally unique" implies that if information is held on an entity at the unclassified, SECRET, and TOP SECRET levels that some process is in place to coordinate the assignment and management of these identifiers across domains. It is not clear how to do so, and there is currently no processes in place to ensure any such uniqueness.

class called `People` that has a property `Weight`, and the units are in pounds, while Data Source$_2$ has a class called `People` that has a property `Weight`, and the units are in kilograms, then harmonization is bringing them into common units, pounds or kilograms (whichever the ontology of the Integrated Entity Knowledge Base uses). A deeper example is if Data Source$_1$ has a property of class `People` that is called `Occupation`, whose values are from the list of occupations from the US Department of Commerce, while Data Source$_2$ has a property of the same class called `Occupation`, but whose values are from the list of occupations from the International Civil Service Commission. It would be necessary to map each of the sets of values into whatever values are defined in the Integrated Entity Knowledge Base's ontology. There might be a simple, one-to-one mapping between the values of the two sets, or the mapping might be quite complicated and not one-to-one (which implies some information may be lost in the mapping), but in any case the mapping would have to be exercised before integration. Otherwise it might appear that two instances of the class `Person` have different occupations when in fact they don't, but they are just called different terms. This harmonization step is particularly important for disambiguation, as described below. This processing must not destroy the connections back to the original entity property value in the Data Source's Entity Knowledge Base, since that is the definitive source for information.

(U) When the entities from the Data Source's Entity Knowledge Bases are harmonized, they need to be stored in the Integrated Entities Knowledge Base. The requirement for such storage is for indexing and thus making it responsive to queries posed by applications. Thus an appropriate search interface (including a query language) must be included as part of the implementation of the Integrated Entities Knowledge Base. There is a strong temptation is to use relational database management (RDBMS) technology as the basis for the storage and indexing, since it is mature and available, and indeed this is what is often done today. For entities and their attributes and relationships this may not be the best approach, due to the relational model that underlies these databases not being able to represent and search entities efficiently[30]. The set of interconnected entities, connected by their relationships, is known as a *semantic graph* (or a semantic web or a semantic network), and a specialized set of databases has arisen to store and index semantic graphs that are of interest to intelligence, which are called *triple stores* (from the triple *object-property-value* used in RDF)[31]. They are optimized for the kinds of

[30] (U) The question often asked is whether or not RDBMS technology **can** work for this type of storage and search. The answer is, "of course they can," but the real question should be the **efficiency** of the approach. Specialized approaches to representing and storing/searching entities are done for efficiency reasons, which, when translated into the billions of entities that are needed to be processed in the intelligence problem, are very important to consider. From a recent Microsoft document: "There are two main benefits offered by a profile store that has been created by using RDF. The first is that RDF enables you to store data in a flexible schema so you can store additional types of information that you might have been unaware of when you originally designed the schema. The second is that it helps you to create Web-like relationships between data, which is not easily done in a typical relational database." See http://msdn2.microsoft.com/en-us/library/aa303446.aspx.

[31] (U) This terminology is not uniformly used. They are also called knowledge bases, object bases, etc. Furthermore, there are other approaches that use variations on the *object-property-value* triple model. In order to not discuss this issue in such broad generalities that little can be said, in this report we will assume that some variation of the *object-property-value* model is used.

search that are of interest to semantic graphs. Many of these triple stores are based on the work done at the W3C—the World Wide Web Consortium—that has standardized on a set of languages to represent and query semantic graphs. The language for representation is based on XML, but adds the ability to express some semantics (meaning) to the tags that XML allows. These languages are called RDF = Resource Description Framework and OWL = Ontology Web Language, with its attendant query language SPARQL[32]. These are not the only languages available for representing semantic graphs (Common Logic, for example, is another option[33]), but we will use these in the discussion in this report.

(U) An implementation of an Integrated Entities Knowledge Base will be done in some specific triple store, which implies a representation language for the entities, the language supported by the triple store. Although the language may be specified by the specific triple store, the ontology by which the "things of interest" are described needs to be known by the triple store. That is, the hierarchy of the classes of which entities can be an instance, along with the properties of each class, must be specified, and in a form that is understandable to the triple store (that is, in the language of the triple store, generally OWL). The properties of each class are inherited from the parent class, although overriding is possible. Two kinds of entity properties should be able to be expressed and stored, *datatype properties*, whose values are numbers, strings, etc., and *object properties*, whose values are other entities. It is the object properties that connect entities to others in the semantic graph. The examples of harmonization above mainly were about datatype properties, but a much more important harmonization will be on object properties, since this is where much of the "meat" of a problem will be.

(U) Once the entities are brought into semantic harmony and stored, they can be integrated. By this we mean that the property values can be combined. For example, if Data Source$_1$ has an entity named `JohnSmith` that has `Weight` 200, and Data Source$_2$ has an entity named `JohnSmith` that has `Weight` 89, with the first in pounds and the second in kilograms, and we harmonize into values 200 and 196 pounds, then integrating them might be to average the values into a single value, 198. But it is not so clear what to do about non-numerical values. For example, what if Data Source$_1$ has an entity named `JohnSmith` that has `ColorHair` Red, and Data Source$_2$ has an entity named `JohnSmith` that has `ColorHair` Auburn. What is the "average" of red hair and auburn hair? Even for numerical values, problems can arise. For example, what if Data Source$_1$ has an entity named `JohnSmith` that has `MeetingAttended` with `Date` 8 July, and Data Source$_2$ has an entity named `JohnSmith` that has `MeetingAttended` with `Date` 8-10 July. In this case, what number should be used as the combination? One might argue that combining values is not necessary, and that the Integrated Entity Knowledge Base should store all values (with a pointer back to the entity in the original Data Source Entity Knowledge Base). But then we still will run into problems of querying the data in the Integrated Entity

[32] (U) For a description and specification of these languages, see http://www.w3.org/RDF/, http://www.w3.org/OWL/, and http://www.w3.org/2001/sw/DataAccess/, respectively.

[33] (U) See http://common-logic.org/.

Knowledge Base for analysis and for presentation to users. It seems like some sort of integration is necessary.

(U) An important aspect of an Integrated Entities Knowledge Base is pedigree and lineage[34]. Since the purpose of the storage of entities is to perform intelligence analysis on the entities, the veracity of the property values must be able to be inferred. To do this usually means that the source of the information and the processing steps it has gone through (and by whom) are critical to the analysis. It is very important, given that the use of the data in the Integrated Entities Knowledge Base is for intelligence purposes, to have a good approach to capturing and using the pedigree and lineage of property values. Basically, the ontology must include class(es) for this purpose, and the appropriate values must be captured, stored, and processed when the entity property is accessed.

(U) In implementing the Integrated Entities Knowledge Base, there will be significant issues of centralization vs. federation. Again it is tempting to take all the entities from all Entities Knowledge Bases and index it in one application (centralization) that is searchable, but this is both technically and organizationally impractical. Rather, some storage and indexing will be done in its own Entities Knowledge Base to serve the local needs of the organization, and no doubt all local processing will not be in the same storage approach, with the same query language, with the same kinds of results, etc. So in an Integrated Entities Knowledge Base there will need to be some kind of federation among these separate Entities Knowledge Bases, a technically challenging problem. The usual approach to this problem is some variation of *brokering and mediation*. Brokering is the process by which a decision is made as to what Entities Knowledge Bases to search, so that queries are not issued to those that are unlikely to contain meaningful results (otherwise the Entities Knowledge Bases may be overloaded processing queries that have a high probability of returning nothing). Mediation is the process by which a query in a common form is translated into a form that the individual Entities Knowledge Bases can process, both the syntax and semantics of the query[35], and the process by which the results of a query are translated from that of the Entities Knowledge Bases into the common form of the Integrated Entities Knowledge Base. Both of these translation steps are potentially difficult, and many issues arise, such as how to perform query relaxation and how to combine relevance ranked resources when the ranking algorithms are not commensurate. However, these processes are vital to giving user applications the view that all the entity data they need is available and searchable.

(U) As shown in the Figure, once harmonization and integration are done, the entities are stored in the Integrated Entities Knowledge Base. Then, Entity Metadata Integrity enforcement may be done. There are many kinds of enforcement that might be done to ensure quality of the data. One especially important integrity processing to the analysis of intelligence is *disambiguation* processing (also called co-reference resolution). This processing is to find multiple entities in the Integrated Entities Knowledge Base that

[34] (U) These terms are not uniformly agreed upon. Pedigree and lineage usually are defined as the list of sources for a property value, keeping track of all the original intelligence that contributes to a value. Other terms used for this concept include provenance and source reference.

[35] (U) Translating syntax tends to be easy; translating semantics can be from hard to very hard to impossible.

actually refer to the same entity in the real world—the same person, place, organization, etc. When two or more entities are determined to be the same, they may be combined into a single entity. This is important since the intelligence analysis potentially needs to use "all that is known" about an entity, which requires combining multiple entities into a single entity in the Integrated Entities Knowledge Base.

(U) Disambiguation processing can be difficult to design, and very complex to implement. Many approaches that have been taken that rely on the name of an entity. That is, if one entity is named `WilliamWilson` and another is named `BillWilson`, we might conclude that they are the same person in the real world[36]. Common sense tells us that this is insufficient, since there could easily be many people in the Integrated Entities Knowledge Base with that name. Also, this approach will not work well for people's names in certain cultures, that do not have a simple relationship between different names to which a person may be referred. More sophisticated approaches recognize that a person's name is only one property that can be used for disambiguation. If, for example, we knew that `WilliamWilson` lived in Peoria, IL and `BillWilson` lived in Seattle, WA, then we probably would not assume they are the same person[37]. In general, good disambiguation processing takes into account all the properties of an entity, both datatype and object. How to decide if two property values are "close enough" is especially difficult in the case of object properties, which have values that are other entities that themselves might not be disambiguated with other entities.

(U) When it comes to implementation of disambiguation approaches, two main factors come into play. The first is the algorithm for deciding if two entities are indeed referring to the same entity in the real world, as discussed above. The second factor is how to enumerate through the entities to compare them for potential disambiguation. The naïve approach simply starts with an entity, compares it to all others, and then goes to the next. The problem with this approach is that it requires $N^2/2$ compares, if N is the number of entities in the graph. For large graphs, this is too computationally intensive[38]. Implementation approaches need to recognize the processing time issues in developing disambiguation processing. In addition, there is an issue of whether the entities that are decided are referring to the same real-world entity are actually combined in the knowledge base, or if there is just a link between them saying that they are the same real-world entity (in OWL, there is a construct called `SameAs` that accomplished this declaration). The former improves access processing performance, but if this is done it is difficult, and maybe impossible, to break them apart later if new data indicates that the

[36] (U) This may seem obvious, but the processing would have to know that "Bill" is a common nickname for "William," a fact that Americans would know but a processing algorithm won't, unless it is "told" in the appropriate form to use for processing.

[37] (U) There is a temporal aspect to this data, so if `WilliamWilson` lived in Peoria in 1998 and `BillWilson` lived in Seattle in 2007, they might in fact be the same person. This temporal nature of data further complicates the disambiguation processing.

[38] (U) Its actually worse, for if two entities are combined, they should be compared to all others again as a combined entity. It is also possible that combining two entities might cause some other two entities to combine, if the first combined entity is a value of some property of one of the latter two, so then each combination decision must be followed by comparing all existing entities, which is order N^3.

two should not have been combined. Thus the latter approach seems best unless there is very high confidence in the combination decision.

(U) Processing entities

(U) The Integrated Entities Knowledge Base will be used by intelligence analysts to support their analyses, resulting in higher quality analysis than available today[39]. There needs to be a query interface that allows analysts to search the Integrated Entities Knowledge Base for entities of interest, and there needs to be analysis tools, such as visualization or link analysis tools, that interface to the Integrated Entities Knowledge Base, as shown in the figure. In both cases the underlying integrated entities with their properties will provide significantly better data against which to operate, by virtue of the semantic integration and disambiguation of entity data from many Data Sources in a common system.

(U) Querying an Integrated Entities Knowledge Base is not the same as querying a relational database, since the entities and their relationships have richer content and are organized in a more natural way for this kind of information, namely a semantic graph. Thus the query interface will need to be user friendly and provide the functionality to allow an analyst to fully explore the set of integrated entities. This capability is particularly important, since if all the data in the world is available but user applications can't effectively find the relevant data[40] for a specific analysis, then any downstream processing might very well work poorly at best. Note that search should be supported both in retrospective mode as well as profiling mode, so that standing queries can be established and results sent to users or their applications upon receipt of entity property values that matches the criteria of the standing query.

(U) There will be times when a query will result in a small number of entities, which the user then can view individually. How to present an entity is not obvious, since the values of many properties of an entity are other entities. For example, the query might ask for all that is known about a specific person (technically, an instance of the class `Person`), and some of the properties of this person are familial relationships, like `CousinOf`. Say one of the values of `CousinOf` is another entity in the Integrated Entities Knowledge Base of class *person* whose name is unknown, but he is known to have participated in a specific event, which is itself another entity in the Integrated Entities Knowledge Base of class *event*. How do we present this to the user? In general, it will be straightforward to present datatype property values, but not object property values, for it will not be obvious how "much" of the entity that is the value of the property to present. We cannot simply say that we will present all entities that are the values of object properties of the entity of interest, for they may only make sense in terms of other entities, and we may end up presenting the entire knowledge base! This is an example of what has become known as

[39] (U) Today analysts are forced to read resources, extract entities and their attributes and relationships manually, keep this data in some local form such as a spreadsheet or Analyst Notebook diagram, and manually integrate across differing Data Sources.

[40] (U) In the sense of precision and recall.

the "six degrees of Kevin Bacon" problem[41]. The most successful presentation methods seem to allow an analyst to expand or contract the depth of the entities connected to the entity returned from the query. As can be seen, there is not a well-accepted way to even perform the simple task of viewing a known entity.

(U) Although sometimes a single, known entity returned from a query will satisfy the needs of an analyst, this will not always be the case. More often, a query will result in a large number of entities, and there needs to be methods and tools to facilitate visualization and analysis of this set of entities. Such visualizations can include timeline displays or geographic displays of the entities, thus helping the analyst understand the set as a whole. Another analyst capability would be successive refinement of queries (called "faceted search"), a process that helps an analyst make good queries by providing feedback on the makeup of a set of entities derived from a broad query, so he or she can see explicitly how to refine the query. A significant issue with presenting results is that of pedigree and lineage. How is this information included in a display of results, so that an analyst gets some sense for how much he or she should trust the information presented.

(U) One particularly significant analysis that will be done on the entities is the identification and classification of patterns of interest in the data. Patterns are partially (or fully) uninstantiated sets of entities and properties, and can be models of behavior of interest (like behavior leading up to a terrorist attack on a certain type of asset). Searching for patterns is a very important use of an Integrated Entities Knowledge Base, so tools will support expression of patterns, search for patterns, analysis of results, and presentation of results.

(U) Lastly, the entities in the Integrated Entities Knowledge Base will be useful to "inform" other applications, where the analyst doesn't even know that he or she is accessing these entities. For example, if a wiki is being used as a collaboration tool for analysis, and a mention is made of a particular person, for example, there may be a link from that mention to the entity in the underlying Integrated Entities Knowledge Base. When this link is clicked, a dynamic web page is created that presents what is known about this entity in some form, with some navigation method for further exploring the entities. Although this is a query to the Integrated Entities Knowledge Base, the analyst does not get this sense of the interaction; he or she just gets what is known throughout the IC on the entity. One advantage to this kind of approach is that, using pedigree and lineage, we can either present what is currently known about the entity or what was known at the time of the assertion in the wiki. Another use of the entity data is by authors of intelligence reports; they can access the GUIDE during the authoring process to include in their products, thus reducing ambiguity in the interpretation of the information in their report.

(U) The previous processing of the integrated entities was by an analyst or application operating directly against the Integrated Entities Knowledge Base. Another very important use of the integrated entities is to provide information back to the Data Sources that provided it. Two types of information may be provided back: property values and

[41] (U) See http://en.wikipedia.org/wiki/Six_degrees_of_separation.

disambiguation information. The former type of information will enrich the Data Source's own Entity Knowledge Base with additional property values that are derived from other Data Sources input resources. In order to properly use these values, the pedigree and lineage must be handed back to the original Data Source along with property values, if it can be (security considerations might prohibit it). Both the property values and the pedigree and lineage must be in a form understandable to the original Data Source. That is, just as the data from the Data Source's Entities Knowledge Base must be translated into the semantics of the Integrated Entities Knowledge Base (the process we called *semantic harmonization*) to be able to be stored and processed, any data from the Integrated Entities Knowledge Base must be translated into the semantics of the original Data Source's Entities Knowledge Base to be able to be stored and processed. This "reverse harmonization" is similar to the harmonization done to integrate, and as such it is quite probably lossy. When developing the mappings from the Data Sources' ontologies into the Integrated Entities Knowledge Base's ontologies, the reverse mappings should also be developed so this step is facilitated. The losses in information when mapped from the original ontologies to the common ontology and back should be minimized. But, in general, information will be lost, and more is likely to be lost in the reverse harmonization since the original ontologies will probably be less complete than the common ontology. An example of the kind of loss that can happen is if the Integrated Entities Knowledge Base has location in latitude and longitude, while the Data Source only has placename. Then, reverse harmonization of a particular lat/lon will result in the city or town where the lat/lon is, but since this is coarser granularity than the lat/lon, information is lost. If the placename were to be contributed back to the Integrated Entities Knowledge Base, say by giving the lat/lon of the city center, then it is clear that information was lost.

(U) Notice that one type of loss will be if the Data Source does not have a property defined in which to store a value. For example, say Data Source$_1$ has in its Entity Knowledge Base a class called `Person` with properties `PassportNumber`, `Address`, `Height`, and `Weight`, while Data Source$_2$ has in its Entity Knowledge Base a class called `Person` with properties `PassportNumber`, `Address`, and `Age`. When combined in the Integrated Entity Knowledge Base, the class called `Person` has properties `PassportNumber`, `Address`, `Height`, `Weight`, and `Age`. Let's say that both Data Sources have information on the same Person, namely `JoeBlow`, who, according to Data Source$_1$ has `PassportNumber` = `123456`, `Address` = `23 Main, Peoria, IL, USA`, `Height` = `5′11″`, and `Weight` = `190#`, while according to Data Source$_2$ has PassportNumber = 123456, Address = `23 Main Street, Peoria, IL`, and `Age` = `34`. When combined, `JoeBlow` has `PassportNumber` = `123456`, `Address` = `23 Main Street, Peoria, IL`, `Height` = `5′11″`, `Weight` = `190#`, and `Age` = `34`. If we hand back to Data Source$_1$ that `JoeBlow` has `Age` = `34`, where will it store this information? Its ontology does not have a property called Age (or something semantically similar), so it has no place to store it. In general, a Data Source can only store and utilize property values that are in its own ontology, which may be significantly fewer than in the common, integrated ontology.

(U) The other information that may be returned to a Data Source is disambiguation information. This in fact may be the most valuable contribution that the Integrated Entities Knowledge Base can contribute to each Data Source's Entities Knowledge Base.

The disambiguation done in the Integrated Entities Knowledge Base, by virtue of the increased number of property values (and by increases in the confidence in the values by multiple collections of data that contributes to the values), is likely to be better than any one Data Source can do. Thus the Integrated Entities Knowledge Base's disambiguation's can be handed back to each original Data Source's Entities Knowledge Base, so long as entity identifiers are properly kept (which will be a necessary part of the pedigree and lineage).

(U) As in any system used for intelligence, there is a security overlay that impacts all processing. This aspect has been downplayed in this report, but in handing back information to original Data Sources, it cannot be ignored. One especially intriguing possibility is that the integrated entities can be used for processing, such as disambiguation, where the original Data Sources do not have access to the same level of information. Then, it is possible that the Integrated Entities Knowledge Base disambiguation processing can conclude that two entities are in fact referring to the same real-world entity, and pass this information back to individual Data Sources, but these Data Sources cannot know why this disambiguation decision was reached, since it may involve property values whose pedigree or lineage may reveal sources or methods that are too sensitive. But this does not mean that the disambiguation decision can't be passed back to the original Data Sources, which results in high value use of all data without violating security models.

(U) Appendix C. Commercial Products Reference Data (U)

(U) The following provides the collected reference data on commercial and open source products that have some capability that fits within the Catalyst needs. First is the list of products with the following:

- Whether they are commercial or open source
- The name of the product
- The name of the company that sells the product (or, in the case of open source, the organization that nominally is in charge of the product)
- The url where the product can be found on the Internet
- A short description of the product

The remainder of this appendix provides the products sorted by functional category, with all the other functional categories that each product is in, plus the totals of commercial and open source products in that category. This is done for each of the categories as delineated in Section 4. All of this information is available on a single spreadsheet that is available by contacting the author.

(U) Many resources were used to find and understand these products. Primary among them are the following (in no particular order):

- AI3 Comprehensive Listing of 175 Semantic Web tools http://www.mkbergman.com/?p=287
- Large Triple Stores, predictions of what some software might scale to http://esw.w3.org/topic/LargeTripleStores
- LingPipe Competition, software available for linguistic processing http://alias-i.com/lingpipe/web/competition.html
- SemanticWebTools http://esw.w3.org/topic/SemanticWebTools
- Sentiment Analysis and Language Processing Tools http://lordpimpington.com/codespeaks/drupal-5.1/?q=node/5
- Dot.Kom Information Extraction Tools http://nlp.shef.ac.uk/dot.kom/technology.html
- SemWebCentral http://projects.semwebcentral.org/softwaremap/trove_list.php
- Text Analytics Wiki http://textanalytics.wikidot.com/commercial
- Text mining, Software and applications http://en.wikipedia.org/wiki/Text_mining
- Ontology Editors http://www.xml.com/2002/11/06/Ontology_Editor_Survey.html
- Mills Davis SemanticWeb Report http://www.project10x.com/
- Message Understanding Conference (1997) http://en.wikipedia.org/wiki/Message_Understanding_Conference
- Exploiting Lexical & Encyclopedic Resources for Entity Disambiguation (2007) http://www.clsp.jhu.edu/ws2007/groups/elerfed/
- MUC-6 http://www.clsp.jhu.edu/ws2007/groups/elerfed/documents/Entity-Disambiguation-Scoring-Metrics.v2.ppt

- RDF Scalability
  http://www.ontotext.com/publications/ScalableReasoningTargets_nov07ak.pdf
- Information Extraction Surveys of state of the Industry old (1996)
  http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html
- State of the Industry of Semantic Web (March 7, 2008)
  http://www.net.intap.or.jp/INTAP/s-web/swc2008/1_Karl.pdf

Complete List

ID	Product	URL	Description
1	21st Century Technologies Large Scale Data Searching	http://www.21technologies.com/index.php?option=com_content&task=view&id=11&Itemid=13	21st Century Technologies Large Scale Data Searching provide for large-scale search operations and real-world distastes and analysis in large graph-based data-stores.
2	21st Century Technologies Lynxeon	http://www.21technologies.com/index.php?option=com_content&task=view&id=8&Itemid=13	2st Century Technologies Lynxeon provides a platform and tools for high-performance pattern search, management, and application development. Built to perform rapidly on very large scale datasets (e.g., billions/trillions of data elements).
3	21st Century Technologies Threat Detection and Analysis (TMODS)	http://www.21technologies.com/index.php?option=com_content&task=view&id=10&Itemid=13	2st Century Technologies Threat Detection and Analysis apply advanced techniques in graph analytics, including subgraph isomorphism, social network analysis (SNA), behavioral modeling, and data fusion to discover powerful new ways to perform threat detect
4	3store	http://www.aktors.org/technologies/3store/	3Store is a MySQL based triple store, currently holding over 30 million RDF triples used by a range of Knowledgeable Services developed within the AKT project.
5	AeroText Core Knowledge Base	http://www.lockheedmartin.com/products/AeroText/index.html	The AeroText product suite provides a fast, agile information extraction system for developing knowledge-based content analysis applications. Possible applications include automatic database generation, routing, browsing, summarizing and searching.
6	AeroText Knowledge Base Engine	http://www.lockheedmartin.com/products/AeroText/index.html	AeroText Knowledge Base Engine a data-independent design applies a knowledge base to your documents
7	AllegroGraph	http://agraph.franz.com/allegrograph/	AllegroGraph is a modern, high-performance, persistent, disk-based RDF Graph database for support for SPARQL, RDFS++, and Prolog reasoning from Java applications.
8	Altova Semantic Web Tool	http://www.altova.com/products_semanticworks.html	Altova SemanticWorks is a visual RDF and OWL editor that graphically designs RDF instance documents, RDFS vocabularies, and OWL ontologies.
9	ANNIE	http://www.aktors.org/technologies/annie/	ANNIE is an open-source, robust Information Extraction (IE) system which relies on finite state algorithms. ANNIE consists of the following main language processing tools: tokeniser, sentence splitter, POS tagger, named entity recogniser. ANNIE can be use
10	Apache Agora	http://people.apache.org/~stefano/agora/	Agora is a virtual community visualizer.
11	Arabic Named Entity Extractor (ANEE)	http://www.coltec.net/default.aspx?tabid=221	ANEE provides effective entity extraction application for Arabic data, utilizing a proprietary taxonomy developed by leading Arabic linguistic scientists.

ID	Product	URL	Description
12	Attensity Explore\Analytics	http://www.attensity.com/products/	Attensity's Explore\Analytics provides business users with drill down and visualization tools to slice, dice and analyze important facts and aggregations of facts extracted from text using Attensity's extraction engines.
13	Attensity Extraction Engine	http://www.attensity.com/products/	Attensity's Extraction Engines extract who, what, where, when, and why, and how and allows users to drill down to understand people, places and events and how they are related.
14	Attensity Search	http://www.attensity.com/products/	Attensity's Text Search is a powerful application for searching text documents that ultimately combines text search with Text Analytics.
15	Attensity Solution Processors	http://www.attensity.com/products/	Attensity's Solution Processors take output from the Attensity extraction engines and make that output appropriate for use in other applications and tools.
16	AXIS	http://www.tactical.overwatch.com/products.asp	AXIS specifically creates diagrams consisting of entities and links arranged in a connected graph.
17	Balie (See also YooName)	http://balie.sourceforge.net/	Balie or Baseline Information Extraction is a multiligual information extraction from text with machine learning and natural language techniques.
18	Basic Formal Ontology (BFO)	http://www.ifomis.org/bfo	Basic Formal Ontology (BFO) grows out of a philosophical orientation which overlaps with that of DOLCE and SUMO.
19	BBN Asio Cartographer	http://asio.bbn.com/cartographer.html	Asio Cartographer is a graphical, ontology mapper that is based on the Semantic Web Rule Language (SWRL).
20	BBN Asio Parliament	http://asio.bbn.com/parliament.html	Asio Parliament implements a high-performance storage engine that is compatible with the RDF and OWL standard.
21	BBN Asio Scout	http://asio.bbn.com/scout.html	Asio Scout enables integration of structured data sources.
22	BBN Asio Semantic Query Decomposition	http://asio.bbn.com/parliament.html	The purpose of the Asio Semantic Query Decomposition (SQD) module is to divide a SPARQL query, posed in this unified vocabulary (called the domain ontology), over multiple data sources.
23	BBN Identifinder	http://www.bbn.com/solutions_and_technologies/data_indexing_and_mining/identifinder	BBN IdentiFinder rapidly analyzes electronically-stored text to locate names of corporations, organizations, people, and places, including variations in names.
24	BBN Semantic Bridge for Relational Databases	http://asio.bbn.com/sbrd.html	Asio SBRD's integrates a relational database into our Semantic Distributed Query architecture.
25	BBN Semantic Bridge for Web Services	http://asio.bbn.com/sbws.html	The Asio Semantic Bridge for Web Services (SBWS) is a standalone tool that enables the integration of SOAP-based web services into a Semantic Web environment.
26	Bobcat	http://bobcatonline.com/services.html	BOBCAT provides the capabilities to automatically identify themes of activities, highlight relationships between entities, group entities that are coordinating, visualize relationships spatially, export results to Microsoft Office.

ID	Product	URL	Description
27	Brahms	http://lsdis.cs.uga.edu/projects/semdis/brahms/	Brahms is a fast main-memory RDF/S storage, capable of storing, accessing and querying large ontologies. Idea for system like BRAHMS came after testing other RDF/S storages (like Jena, Sesame or Redland) while using model in main memory.
28	BullDoc	http://www.trifeed.com/product-BULLDOC.htm	BullDoc server will crawl your organization resources (shared directories, submitted emails, specific web sites), feed them to the information extraction engine that will save the extracted data into the database.
29	BusinessObjects Text Analysis	http://www.businessobjects.com/products/platform/textanalysis/	BusinessObjects Text Analysis "reads" text in 30+ languages, extracting key information so unstructured text data can be used as a data source for data integration or business intelligence, uncovering hidden information in CRM systems, Web and e-mails.
30	Carabao DeepAnalyzer	http://www.digitalsonata.com/download.aspx?type=desktop	Carabao DeepAnalyzer lets you search your data for inflections and synonyms of a search argument, search for concepts, and find places, names, phone numbers, medications, weapons, chemical compounds, financial terms, diseases, and more.
31	Carabao Standard Free Edition	http://www.digitalsonata.com/download.aspx?type=desktop	Carabao Standard Edition -Free Includes lexicon development, management and testing tools, and the transliteration console.
32	Centrifuge	http://www.tildenwoods.com/products.html	Centrifuge lets users ask open-ended questions of their data by interacting with visual representations directly.
33	Ceryph Insight	http://www.ceryph.com/	Ceryph is the commercial version of CmapTools and empowers users to construct, navigate, share and criticize knowledge models represented as concept maps.
34	CIA World FactBook	https://www.cia.gov/library/publications/the-world-factbook/index.html	The World Factbook provides national-level information on countries, territories, and dependencies.
35	Cicero	http://www.languagecomputer.com/solutions/information_extraction/cicero/index.html	The Cicero information Extraction Solution scans all documents and extracts all instances that match that information request.
36	CiceroLite	http://www.languagecomputer.com/solutions/information_extraction/cicero_lite/index.html	Cicero Lite enables fast and robust disambiguation of a large category of names, ranging from company names to product names, names of diseases or drugs, biological and biochemical names, e.g. plants, scientific names of genes or chemical compounds.
37	Clarabridge Business Intelligence Search	http://www.clarabridge.com/Products/BISearch/tabid/106/Default.aspx	Clarabridge BI Search allows business users to easily query existing reports through a Google-like interface, greatly improving their ability to gain insight from your existing business intelligence content (e.g., reports, metrics, and analytics).
38	Clarabridge Content Mining Platform	http://www.clarabridge.com/Products/ContentMiningPlatform/tabid/105/Default.aspx	The Clarabridge Content Mining platform delivers the unstructured content (e-mail, blogs, chat session) into whichever analytical tool the end user feels is appropriate to the task at hand, including business intelligence, data mining and visualization.

ID	Product	URL	Description
39	Classifier4J	http://classifier4j.sourceforge.net/index.html	Classifier4J is a Java library designed to do text classification. It comes with an implementation of a Bayesian classifier, and now has some other features, including a text summary facility.
40	ClearForest Analytics	http://www.clearforest.com/Technology/Tags.asp	With ClearForest text analytics, organizations can systematically incorporate text into their business intelligence systems. It is designed to help analysts and researchers quickly visualize complex associations, relationships and concepts
41	ClearForest Extraction Modules	http://www.clearforest.com/Technology/TechnologyOverview.asp	ClearForest's advanced text-driven business intelligence solutions apply intelligent mark-up to key entities such as person, organization, location, as well as detailed facts or events embedded within free-form text such as news articles and web surveys.
42	COGITO Discover	http://www.expertsystem.net/page.asp?id=1521&idd=27	COGITO Discover is the activity which allows the extraction, transformation and loading of data."
43	COGITO Intelligence	http://www.expertsystem.net/page.asp?id=1521&idd=25	COGITO Intelligence traces all information, identifies the structural and lexical aspects of a text, identifies the conceptual links between various documents and carries out disambiguation and advanced semantic comprehensions operations.
44	COGITO Semantic Search	http://www.expertsystem.net/page.asp?id=1521&idd=18	By leveraging computational linguistic tools, Expert System's Semantic Technology enables the creation of knowledge from the management of information extracted from different kinds of documents.
45	Common Terrorism Information Sharing Standards (CTISS)	http://www.ise.gov/pages/ctiss.html	CTISS program allows for business process-driven, performance-based "common standards" for preparing terrorism information for maximum distribution and access."
46	Connexor Machinese Metadata	http://www.connexor.eu/technology/machinese/machinesemetadata/	Connexor Machinese Metadata is a high-performance text analytics and metadata automation solution, which extracts information, analysts can find hidden story; trends, anomalies, entities.
47	Content Analyst Latent Semantic Indexing	http://contentanalyst.com/html/technologies.html	Latent Semantic Indexing technology is designed to extract every contextual relation among every term in every text object within a collection.
48	CORDER (COmmunity Relation Discovery by named Entity Recognition)	http://kmi.open.ac.uk/projects/corder/	CORDER (COmmunity Relation Discovery by named Entity Recognition) discovers relations from the Web pages of the community.
49	Cyc Knowledge Base	http://www.cyc.com/cyc/company/about	The Cyc software combines an unparalleled common sense ontology and knowledge base with a powerful reasoning engine and natural language interfaces to enable the development of novel knowledge-intensive applications.
50	Cycorp OpenCyc	http://www.cyc.com/cyc/opencyc/overview	OpenCyc is the open source version of the Cyc technology, the world's largest and most complete general knowledge base and commonsense reasoning engine.

ID	Product	URL	Description
51	Cymfony Content Analysis (Info Extact Engine) Engine	http://www.cymfony.com/sol_dash_eng.asp	MI/Cymfony's is an advanced information extraction engine that combines information retrieval and Natural Language Processing (NLP) technologies to identify important people, places, companies, concepts, relationships and events in documents.
52	Cymfony Orchestra	http://www.cymfony.com/sol_orchestra.asp	TNS MI/Cymfony's Orchestra enables clients to see emerging trends, product problems and service issues relevant to your company, products and competitors.
53	D2R Server	http://sites.wiwiss.fu-berlin.de/suhl/bizer/d2r-server/	D2R Server, turns relational databases into SPARQL endpoints, based on Jena's Joseki.
54	DERI Ontology Management Environment (DOME)	http://dome.sourceforge.net/	The DERI Ontology Management Environment (DOME) is developed by the Ontology Management Working Group (OMWG).
55	Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE)	http://www.loa-cnr.it/DOLCE.html	Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) is the foundational ontology for comparing the relationships with other future modules of the library.
56	DIANE Core Server	http://www.precipia.com/diane05.asp	DIANE (Digital Analysis Environment) Core Server manages information throughout the consumption process (collection, organization, visualization, discovery, analysis, and reporting).
57	DIANE Knowledge Services	http://www.precipia.com/diane05.asp	DIANE (Digital Analysis Environment) knowledge services and tools consist of a set of Natural Language Processing engines enabling rapid organization, visualization, and preliminary analysis of unstructured or qualitative data.
58	Digital Reasoning GeoLocator	http://www.digitalreasoning.com/GeoLocator	GeoLocator from Digital Reasoning is a precision-based tool that will extract countries and populated places from unstructured text, while providing their respective geo-coordinates.
59	Digital Reasoning Interceptor	http://www.digitalreasoning.com/Interceptor	Interceptor allows you the ability to look through all of your data rapidly and easily discover what is inside.
60	Dome	http://www.aktors.org/technologies/dome/	A programmable XML editor which is being used in a knowledge extraction role to transform Web pages into RDF, and available as Eclipse plug-ins. DOME stands for DERI Ontology Management Environment.
61	ELIE	http://www.aidanf.net/software/elie_an_adaptive_information_extraction_system	ELIE is a tool for adaptive information extraction from text for Python. It also provides a number of other text processing tools e.g. POS tagging, chunking, gazetteer, stemming.
62	Endeca Information Access Platform	http://endeca.com/technology/index.html	The Endeca Information Access Platform and the MDEX Database Engine helps you successfully build tailored applications for people to explore your existing data, regardless of its source or format.
63	Espotter	http://kmi.open.ac.uk/projects/espotter/	Adaptive Named Entity Recognition for Web Browsing

ID	Product	URL	Description
64	Factiva Taxonomy Warehouse	http://www.taxonomywarehouse.com/	Factiva's Taxonomy Warehouse offers more than 550 taxonomies, arranged in 73 subject domains, produced by 260 publishers in 39 languages. More than 100 of these taxonomies can be licensed directly through Taxonomy Warehouse.
65	FASTUS	http://www.ai.sri.com/~appelt/fastus.html	FASTUS is a (slightly permuted) acronym for Finite State Automa-based Text Understanding System. It is a system for extracting information from free text.
66	Fetch Agent Platform	http://www.fetch.com/products.asp	Fetch Technologies has developed a powerful platform for extracting and integrating information from multiple Web sources, and transforming the data into a form that is useful for business applications.
67	FMS Sentinel Visualizer	http://www.fmsasg.com/Products/SentinelTMS/	Sentinel Visualizer provides data visualization, link analysis, and social network analysis.
68	FreeLing	http://garraf.epsevg.upc.es/freeling/	The FreeLing package consists of a library providing language analysis services.
69	General Architecture for Text Engineering (GATE)	http://www.aktors.org/technologies/gate/	GATE is a stable, robust, and scalable open-source infrastructure which allows users to build and customise language processing components, while it handles mundane tasks like data storage, format analysis and data visualisation.
70	Graphl	http://home.subnet.at/flo/mv/graphl/	Graphl is an RDF tool for collaborative editing and visualisation of graphs, representing relationships between resources or concepts of the real world.
71	Graphviz	http://www.graphviz.org/	Graph visualization is a way of representing structural information as diagrams of abstract graphs and networks.
72	GrOwl	http://ecoinformatics.uvm.edu/technologies/growl-knowledge-modeler.html	GrOWL provides a graphical browser and an editor of OWL ontologies that can be used stand-alone or embedded in a web browser.
73	Guess, The Graph Exploration System	http://graphexploration.cond.org/index.html	GUESS is an exploratory data analysis and visualization tool for graphs and networks.
74	Hozo Ontology Editor	http://www.hozo.jp/	An environment for building\using ontologies.
75	HP Labs Jena	http://jena.sourceforge.net/index.html	Jena is a Java framework for building Semantic Web applications. It provides a programmatic environment for RDF, RDFS and OWL, SPARQL and includes a rule-based inference engine.
76	HP Labs SDB	http://jena.hpl.hp.com/wiki/SDB	SDB is a component of Jena for the RDF storage and query specifically to support SPARQL.
77	HyperTree Java Library	http://hypertree.sourceforge.net/	An hyperbolic tree visualization java library, to implement hyperbolic tree easily. See http://www.inxight.com for explanations and examples.
78	i2 Analyst's Notebook	http://www.i2.co.uk/Products/Analysts_Notebook/default.asp	Analyst's Notebook provides the optimum environment for effective link and timeline analysis.
79	i2 ChartExplorer	http://www.i2.co.uk/products/i2chartexplorer/	i2 ChartExplorer It allows you to find, explore and re-use information in charts and documents stored in

ID	Product	URL	Description
			PCs and servers on your network.
80	IBM Entity Analytic Solutions	http://www-306.ibm.com/software/data/ips/products/masterdata/eas/	EAS provides real time identity and relationship recognition and resolution in context with business applications.
81	IBM Global Name Recognition	http://www-306.ibm.com/software/data/ips/products/masterdata/globalname/	IBM Global Name Recognition products lead in providing multi-cultural name recognition software solutions for mission critical applications.
82	IBM Information Server	http://www-306.ibm.com/software/data/integration/info_server_platform/	IBM Information Server is a revolutionary new data integration software platform from IBM that helps organizations derive more value from the complex, heterogeneous information spread across their systems.
83	IBM Integrated Ontology Development Toolkit (IODT)	http://www.alphaworks.ibm.com/tech/semanticstk	IODT is a toolkit for ontology-driven development.
84	IBM Multiplatform Master Data Management	http://www-306.ibm.com/software/data/ips/products/masterdata/	IBM Multiform Master Data Management manages master data domains (customers, accounts, products) that have a significant impact on the most important business processes and realizes the promise of SOA.
85	IBM Semantic Layered Research Program (Boca)	http://ibm-slrp.sourceforge.net/	Boca system is a server capable of storing millions of RDF triples in a DB2 database.
86	IBM Web Ontology Manager	http://www.alphaworks.ibm.com/tech/wom?open&S_TACT=105AGX59&S_CMP=GR&ca=dgr-lnxwd01awwom	A Web-based system for managing Web Ontology Language (OWL) ontologies.
87	IHMC CmapTools	http://cmap.ihmc.us/download/free_client.php?myPlat=Win	The CmapTools program empowers users to construct, navigate, share and criticize knowledge models represented as concept maps.
88	Inflow	http://www.orgnet.com/inflow3.html	Orgnet.com provides social network analysis software & services for organizations, communities, and their consultants.
89	Infogistics Xtractor	http://www.infogistics.com/xtractor.html	Xtractor is an engine that sifts through large volumes of texts and creates database records for the objects that are mentioned in the text, such as people, organisations, locations, vehicles, etc.
90	Initiate Customer	http://www.initiatesystems.com/products_services/mds/consumer/Pages/default.aspx	Initiate Consumer enables you to know your customer with confidence, whenever and wherever that customer is encountered.
91	Initiate Master Data Service	http://www.initiatesystems.com/products_services/mds/Pages/default.aspx	Initiate software provides organizations with complete, highly accurate and real-time views of data spread across mulitple systems or databases.
92	Initiate Organization	http://www.initiatesystems.com/products_services/mds/organization/Pages/default.a	Initiate Organization brings together customer and organizational hierarchy data from multiple sources to provide a comprehensive view of each customer and

ID	Product	URL	Description
		spx	how it fits into a larger context.
93	Insightful InFact (Evri Solutions)	http://www.evri.com/	Insightful Miner is a powerful, scalable, data mining and analysis workbench that enables organizations to deliver customized predictive intelligence where and how it is needed.
94	Insightful Miner	http://www.insightful.com/products/iminer/default.asp	Insightful Miner is a powerful, scalable, data mining and analysis workbench that enables organizations to deliver customized predictive intelligence where and how it is needed.
95	Intellidimension InferEd	http://www.intellidimension.com/pages/site/products/infered/default.rsp	InferEd is an authoring environment to navigate and edit RDF.
96	Intellidimension RDF Gateway	http://www.intellidimension.com/pages/site/products/rdfgateway.rsp	RDF Gateway is a high-performance, scalable semantic web server with a RDF deductive database at its core.
97	Intelligenxia uReveal	http://www.intelligenxia.com/Products/uReveal.htm	uReveal pantent analytics for idea extraction and relationship discovery, integrated chart/graphing capabilities.
98	Interwoven MetaTagger	http://www.interwoven.com/components/page.jsp?topic=PRODUCT::METATAGGER	Interwoven MetaTagger automates complex tasks such as creating taxonomy driven Website navigation and tagging content for dynamic presentation. MetaTagger intelligently and automatically categorizes content and extracts information based on business requirements.
99	Interwoven Universal Search	http://www.interwoven.com/components/page.jsp?topic=PRODUCT::UNIVERSAL_SEARCH	Interwoeven Universal Search helps unify content across multiple internal and external content sources within a single search environment.
100	Inxight Metadata Management System	http://www.inxight.com/products/mms/	The Inxight SmartDiscovery Metadata Management System (MMS) allows users to review, cleanse and augment automatically extracted text about entities, relations and events.
101	Inxight Search Extender for Google Desktop	http://www.inxight.com/products/se_google/download.php	Inxight Search Extender for Google Desktop is a stand-alone product that that extends Google Desktop to "go the extra mile" helping you find documents faster and locate hidden information that would otherwise be overlooked.
102	Inxight SmartDiscovery Awareness Server	http://www.inxight.com/products/smartdiscovery_as/	Inxight SmartDiscovery Awarness Server is a federated search solution that finds disparate information and extracts the data with a "human level" understanding of the content.
103	Inxight SmartDiscovery Extraction Server (aka Analysis Server)	http://www.inxight.com/products/sd_es/	Inxight SmartDiscovery Extraction Server (aka Analysis Server) extracts information in 30 languages and comprehensive set of advanced text analysis tools include entity, event and relationship extraction, categorization and summarization.
104	Inxight SmartDiscovery VizServer	http://www.inxight.com/products/vizserver/	Inxight SmartDiscovery VizServer helps you gain that advantage by giving you the ability to dynamically explore relationships, trends and timelines.

ID	Product	URL	Description
105	Inxight ThingFinder SDK	http://www.inxight.com/products/sdks/tf/	The Inxight ThingFinder SDK (Software Development Kit) provides advanced text analysis technology that automatically identifies and extracts key entities or other "things" from any text data source, in multiple languages, with no setup or manual creation.
106	IsaViz: A Visual Authoring Tool for RDF	http://www.w3.org/2001/11/IsaViz/	IsaViz is a visual environment for browsing and authoring RDF models represented as graphs.
107	Janya Semantex	http://www.janyainc.com/products/products_semantex.php	Semantex is an enterprise-class information extraction system that supports the automatic or semi-automatic analysis of large volumes of electronic information in order to detect entities, attributes, relationships and events.
108	Jena with PostgreSQL	http://jena.sourceforge.net/DB/postgresql-howto.html	Jena support for PostgreSQL (pronounced Post-Gres-Q-L.) is an enhancement of the POSTGRES database management system, a DBMS research prototype developed at the University of California-Berkeley in the 1990s.
109	jInFil	http://tcc.itc.it/research/textec/tools-resources/jinfil.html	jInFil is an open source Java tool for Instance Filtering. Instance Filtering is a preprocessing step for supervised classification-based learning systems for entity recognition.
110	Joseki	http://www.joseki.org/	Jena's Joseki layer offers an RDF Triple Store facility with SPARQL interface (see also Jena).
111	Kaidara Text2data	http://www.kaidara.com/Products/P_Modules.htm#Text2Data	Kaidara Text2Data is a tool used to index and transform materials in disparate and unstructured form for inclusion in a knowledgebase.
112	Kofax Capture	http://www.kofax.com/products/ascent/capture/index.asp	Kofax Capture automates information capture from scanned paper or imported electronic documents. Based on criteria you define, the entire document or extracted data is digitized, then routed to an archive, database, or the next step in your business works.
113	Kowari	http://www.kowari.org/	Kowari is an Open Source, massively scalable, transaction-safe, purpose-built database for the storage, retrieval and analysis of metadata.
114	Lexalytics Salience Engine	http://www.lexalytics.com/index-4.html	It provides the low level text analytics capabilities of the system, including: Entity Extraction: identifying People, Companies, Places, Products, Email, and Dates.
115	Leximancer Professional	http://www.leximancer.com/cms//index.php?option=com_content&task=view&id=45&Itemid=86	Leximancer is a software tool that enables users to find meaning from text-based documents. It automatically identifies key themes, concepts and ideas from unstructured text with little or no guidance. The innovative concept map allows users to interact
116	Leximancer Server	http://www.leximancer.com/cms//index.php?option=com_content&task=view&id=47&Itemid=65	Leximancer Server is targeted at enterprise deployments of Leximancer.
117	Lexiquest Mine	http://www.spss.com/lexiquest/lexiquest_mine.htm	With LexiQuest Mine, your organization's analysts and business users can uncover concepts contained in text and see them displayed in a color-coded graphical map.

ID	Product	URL	Description
118	LibSea	http://www.caida.org/tools/visualization/libsea/	LibSea is both a file format and a Java library for representing large directed graphs on disk and in memory. Scalability to graphs with as many as one million nodes has been the primary goal.
119	LingPipe (aka ThreatTrackers)	http://www.alias-i.com/lingpipe/index.html	LingPipe is a suite of Java libraries for the linguistic analysis of human languages.
120	Linguamatics I2E	http://www.linguamatics.com/solutions/ie/solutions_product.html	Linguamatics I2E enables you to answer business-critical questions by rapidly extracting relevant facts and relationships from large document collections.
121	LinKFFactory	http://www.landcglobal.com/pages/linkfactory.php	LinkFactory is specifically designed to build ontologies exceeding millions of concepts.
122	Lucid Threat Management System	http://www.dullesresearch.com/lucid/features	Lucid Identify reveals previously unknown illicit networks hiding in plain sight.
123	MarkLogic Server	http://www.marklogic.com/products/ml_server.html	MarkLogic Server is a content integration, discovery, and analysis system that takes full advantage of XML content through the flexibility of XQuery.
124	Megaputer PolyAnalyst	http://www.megaputer.com/polyanalyst.php	Automated keyword extraction, concept correlation and document summarization.
125	Megaputer TextAnalyst	http://www.megaputer.com/textanalyst.php	TextAnalyst provides document summary and navigation.TextAnalyst can provide you with the ability to perform semantic information retrieval or focus your text exploration around a certain subject.
126	Melingo	http://www.melingo.co.il/ab.htm	Melingo is a leader in computerization of Hebrew. The company offers a unique infrastructure that it is able to break down even the most complex Hebrew texts into their true components.
127	Meta Integration Model Bridge (MIMB) "Metadata Integration" Solution	http://www.metaintegration.net/Products/MIMB/	The Meta Integration Model Bridge (MIMB) product provides MITI's metadata movement solution. MIMB users are typically database and software developers who want to move their metadata (models) between various tools from different vendors, across methodologies.
128	MetaCarta Geographic Text Search (GTS)	http://www.metacarta.com/solutions/products/geographic-text-search.html	GTS identifies implied and explicit references to geographic locations within documents, assigns latitude/longitude coordinates to the references, indexes the document, and then enables a search for indexed documents through Graphical User Interfaces
129	MetaCarta GeoTagger	http://www.metacarta.com/solutions/products/geotagger.html	GeoTagger is a production-level geographic entity resolver that parses documents, extracts geographic references within the content, and resolves the geographic meaning intended by the author.
130	Metatomix m3t4.studio Semantic Tookkit	http://www.m3t4.com/semantic.jsp	The Metatomix Semantic Toolkit is a set of Eclipse plugins that allow developers to create and manage ontologies based on the OWL and RDF standards.
131	Metatomix Semantic Platform	http://www.metatomix.com/	The Metatomix Semantic Platform intelligently connects all of your data in real-time and makes it available to any application.

ID	Product	URL	Description
132	Minor Third	http://minorthird.sourceforge.net/	MinorThird is a collection of Java classes for storing text, annotating text, and learning to extract entities and categorize text.
133	MIPT State Department list of Foreign Terrorist Organizations (FTO)	http://www.tkb.org/FTO.jsp	Foreign Terrorist Organizations (FTOs) are foreign organizations that are designated by the Secretary of State in accordance with the Immigration and Nationality Act (INA).
134	MIPT State Department list of selected Other Terrorist Organizations (OTO)	http://www.tkb.org/OtherTerr.jsp	This list includes other selected terrorist groups also deemed of relevance in the global war on terrorism.
135	MIPT State Department Terrorist Exclusion List (TEL)	http://www.tkb.org/TerrExclusion.jsp	The US Patriot Act of 2001 authorized the Secretary of State, with the request of the Attorney General, to designate terrorist organizations for immigration purposes.
136	MIPT Terrorism Knowledge Base (TKB)	http://www.tkb.org/	A comprehensive databank of global terrorist incidents and organizations (includes groups, leaders & members, cases, incidents, and countries/areas; downloads and analytical tools).
137	Model Futures OWL Editor	http://www.modelfutures.com/owl/	Model Futures have developed a free OWL Editor Tool. The editor is tree-based and has a navigator tool for traversing property and class-instance relationships.
138	Modus Operandi Wave	http://www.modusoperandi.com/products.html	ModusOperandi Wave makes use of an ontology (or conceptual model) to unify and resolve semantic conflicts among data sources.
139	Mulgara (see Kowari)	http://mulgara.org/	The Mulgara Semantic Store is an Open Source, massively scalable, transaction-safe, purpose-built database for the storage and retrieval of RDF, written in Java. It is an active fork of Kowari.
140	NameFinder	http://www.apptek.com/products/namefinder.html	AppTek's NameFinder is an advanced technology engine that is used to scan text for proper nouns (such as human names) in various languages--even in writing systems that do not use capitalization.
141	NCTC Worldwide Incidents Tracking System (WITS)	http://wits.nctc.gov/Main.do	The Worldwide Incidents Tracking System is the National Counterterrorism Center's database of terrorist incidents (includes incidents from 1/1/04 through 9/30/07; exportable to XML/XSD and Oracle 10g).
142	NetMiner	http://www.netminer.com/NetMiner/overview_01.jsp	NetMiner allows you to explore your network data visually and interactively, and helps you to detect underlying patterns and structures of the network.
143	NetOwl Extractor	http://www.netowl.com/products/extractor.html	Accurately perform entity extraction from unstructured texts using advanced computational linguistics and natural language processing.
144	NetOwl InstaLink	http://www.netowl.com/products/instalink.html	Accurately perform entity extraction from unstructured texts using advanced computational linguistics and natural language processing.
145	NetOwl TextMiner	http://www.netowl.com/products/textminer.html	SRA's NetOwl TextMiner is a text mining solution that enables users to find, organize, analyze, and mine a large volume of unstructured information.

ID	Product	URL	Description
146	NGA GEOnet Names Server (GNS)	http://earth-info.nga.mil/gns/html/index.html	The Geographic Names Server is the official repository of standard spellings of all foreign place names, sanctioned by the United States Board on Geographic Names.
147	NMARKUP fact-file	http://www.aktors.org/technologies/nmarkup/	NMARKUP (Using GATE process) helps the user build ontologies by detecting nouns in texts and by providing support for the creation of an ontology based on the entities extracted.
148	oBrowse	http://sourceforge.net/projects/obrowse/	oBrowse is a web based ontology browser developed in java.
149	Ontology Works Integrated Ontology Development Environment	http://www.ontologyworks.com/products/iode	Our development environment for producing ontologies (high fidelity domain models) for compilation to our Ontology Works Knowledge Servers.
150	Ontoprise OntoStudio	http://www.ontoprise.de/content/e1171/e1249/index_eng.html	OntoStudio is a professional development environment for modeling ontologies and administrating ontology-based solutions that allows for the integration of multiple heterogeneous data sources.
151	Ontotext BigOWLIM (see also OWLIM) Semantic Repository	http://www.ontotext.com/owlim/	OWLIM is a high-performance semantic repository developed in Java. BIGOWLIM offers non-trivial OWN inference against 1 Billion triples.
152	Ontotext KIM Platform	http://www.ontotext.com/kim/index.html	KIM is a software platform for co-occurrence tracking and ranking of entities, indexing and retrieval.
153	Ontotext ORDI SG	http://www.ontotext.com/kim/index.html	ORDI SG enables enterprise data integration via an RDF-like tripleset model.
154	Ontotext OWLIM (see also BigOWLIM)	http://www.ontotext.com/products/index.html	OWLIM is an industrial-scale semantic database, using Semantic Web standards for inference and integration/consolidation of heterogeneous data.
155	OntoWare Oyster	http://ontoware.org/projects/oyster/	Oyster is a peer to peer system for storing and sharing ontology Metadata.
156	Open Anzo	http://www.openanzo.org/	Anzo is an open source enterprise-featured RDF store and middleware platform capable of storing millions of RDF triples in an underlying relational database.
157	OpenLink Virtuoso Open Source	http://virtuoso.openlinksw.com/dataspace/dav/wiki/Main/VOSIntro	OpenLink Virtuoso is the open source version of its Virtuoso product, includiing WebDAV/web server and SOA functions."
158	OpenLink Virtuoso Universal Server	http://virtuoso.openlinksw.com/	A Cross Platform Universal Server for SQL, XML, RDF Data Management that also includes a powerful Virtual Database Engine, Web Services Deployment Platform, and Web Application Server."
159	OpenNLP	http://opennlp.sourceforge.net/	OpenNLP also hosts a variety of java-based NLP tools which perform sentence detection, tokenization, pos-tagging, chunking and parsing, named-entity detection, and coreference.
160	Oracle 11g (Spatial)	http://www.oracle.com/technology/products/spatial/index.html	Oracle Spatial 11g includes an open, scalable, secure and reliable RDF management platform.

ID	Product	URL	Description
161	Organizational Risk Assessment (ORA)	http://www.casos.cs.cmu.edu/projects/ora/software.php	ORA (Organizational Risk Assessment) is a dynamic network analysis tool that enables the analysis of both standard social network data and meta-network data.
162	OWL Verbalizer	http://attempto.ifi.uzh.ch/site/docs/verbalizing_owl_in_controlled_english.html	OWL Verbalizer converts an OWL RDF/XML to Attempto Controlled English (ACE).
163	OwlSight	http://pellet.owldl.com/ontology-browser/	OwlSight is a lightweight OWL ontology browser that runs in any modern web browser.
164	Pajek	http://vlado.fmf.uni-lj.si/pub/networks/pajek/	Pajek is a program for analysis and visualization of large networks having some ten or hundred of thousands of vertices.
165	Paladin	http://www.metsci.com/about/paladin.html	Paladin is designed to detect threat activities and network anomalies by efficiently searching massive, noisy data that may be unreliable, incomplete and inconsistent.
166	Palantir	http://www.palantirtech.com/products.html	The system integrates with all existing data sources in the enterprise and ultimately serves as an amplifier to an organization's analytical capabilities.
167	Picoolo Toolkit	http://www.cs.umd.edu/hcil/piccolo/	Piccolo is a toolkit that supports the development of 2D structured graphics programs.
168	pOwl	http://sourceforge.net/projects/powl	Powl is web-based ontology authoring and management solution for the Semantic Web.
169	Prefuse (see also SocialAction)	http://prefuse.org/	Prefuse is a set of software tools for creating rich interactive data visualizations.
170	Protégé	http://protege.stanford.edu/	Prot,g, is a free, open source ontology editor and knowledge-base framework.
171	Proximity	http://kdl.cs.umass.edu/software/	Proximity is an open-source system for relational knowledge discovery.
172	RAP NetAPI	http://sites.wiwiss.fu-berlin.de/suhl/bizer/rdfapi/tutorial/netapi.html	The RAP NetAPI is a server for publishing RDF models on the web.
173	RapidMiner (formerly YALE)	http://rapid-i.com/	RapidMiner is an open-source data mining solution that covers a wide range of real-world data mining tasks.
174	RDF Gravity (RDF Graph Visualization Tool)	http://semweb.salzburgresearch.at/apps/rdf-gravity/index.html	RDF Gravity is a tool for visualising RDF/OWL Graphs/ ontologies.
175	RDF2Go	http://ontoworld.org/wiki/RDF2Go	RDF2Go is an abstraction layer over triple (and quad) stores.
176	RDFe	http://infomesh.net/pyrple/rdfe/	RDFe is a schema-aware RDF editor.
177	RDFStore	http://rdfstore.sourceforge.net/	RDFStore is an RDF storage with Perl and C API-s and SPARQL facilities.
178	Readware Information Processor	http://www.readware.com/prods.asp	Readware discovers themes, topics, issues and names of people, places and products.

ID	Product	URL	Description
179	RelationalOWL	https://sourceforge.net/projects/relational-owl	RelationalOWL automatically extracts the semantics of virtually any relational database and transforms this information automatically into RDF/OWL.
180	Revelytix Knoodl	http://knoodl.com/ui/home.html	Knoodl is sort of an ontology editor, registry/repository, and wiki all rolled into an easy to use online application."
181	Revelytix MatchIt	http://revelytix.com/products.htm	MatchIT, a component of the MetaMatrix Semantic Data Services product, provides automated semantic mapping technology to aid domain experts in more quickly reconciling the semantics across a dispersed information environment.
182	Rosette Cross-Language Toolkit	http://www.basistech.com/cross-language-toolkit/	Rosette Cross-Language Toolkit enables English speakers to search documents in foreign languages.
183	Rosette Entity Extractor	http://www.basistech.com/entity-extraction/	Rosette Entity Extractor locates names, places, dates and other words and phrases in multi-lingual text through advanced named entity extraction."
184	Rosette Name Indexer	http://www.basistech.com/name-indexer/	Rosette Name Indexer matches foreign names across writing systems and languages.
185	SaffronScope	http://www.saffrontech.com/saffron-scope.shtml	SaffronScope is a web-based application that allows discovery of entity-to-entity similarities.
186	SaffronWeb	http://www.saffrontech.com/saffron-web.shtml	SaffronWeb is a web-based application for knowledge discovery and sharing to create both end user and data source memories.
187	Sandpiper Visual Ontology Modeler	http://www.sandsoft.com/index.html	The Visual Ontology Modeler is a tool that enables construction of component-based ontologies allowing businesses to unlock new capabilities and functions in current information stores.
188	SAS Enterprise Miner	http://www.sas.com/technologies/analytics/datamining/miner/	SAS Enterprise Miner streamlines the data mining process to create accurate predictive and descriptive models based on analysis of vast amounts of data from across the enterprise.
189	SAS Model Manager	http://www.sas.com/technologies/analytics/modelmanager/manager/index.html	SAS Model Manager is used to create, manage, and deploy life-cycle analytics."
190	SAS Text Miner	http://www.sas.com/technologies/analytics/datamining/textminer/	SAS Text Miner provides a rich suite of tools for discovering and extracting knowledge from text documents.
191	SchemaLogic Enterprise Suite	http://www.schemalogic.com/products/enterprise_suite/	SchemaLogic Enterprise Suite (SES) enables business subject matter experts and IT professionals to define and manage a semantic standard. Software, services, and integration technologies empower companies to capture and manage standard business terminology.
192	Semantic Research Semantica SE	http://www.semanticresearch.com/products/se.php	Semantica SE allows expert knowledge producers and consumers alike to access, learn and benefit from highly interconnected and easily understood contextual knowledge structures.

ID	Product	URL	Description
193	Semantic Web RDF Library for C#/.NET	http://razor.occams.info/code/semweb/	Semantic Web for RDF Library supports persistent storage in MySQL, Postgre, and Sqlite. The library can be used for reading and writing RDF and supports SPARQL.
194	Sesame	http://www.aduna-software.com/technologies/sesame/overview.view	Sesame is a fast and scalable RDF database.
195	Siderean Seamark MAPP	http://www.siderean.com/products_suite.aspx	Seamark MAPP is a metadata processing platform. It is a scalable and extensible metadata-generation system, built on the open-source UIMA framework to harvest metadata from sources such as MS SharePoint, RSS feeds, Web content and various file systems.
196	Siderean Seamark Navigator	http://www.siderean.com/products_suite.aspx	Seamark Navigator is the relational navigation server. It discovers and indexes content, pre-calculates relationships and suggests paths for data exploration.
197	Smartlogic Ontology Manager	http://www.aprsmartlogik.com/index.php/solutions/ontology	Semaphore OM (Ontology Manager) is the taxonomy and ontology authoring component of the Semaphore Semantic Middleware platform.
198	Snoogle	http://snoggle.projects.semwebcentral.org/	Snoggle is a graphical, SWRL-based ontology mapper to assist in the task of OWL ontology alignment. It allows users to visualize ontologies and then draw mappings from one to another on a graphical canvass.
199	SocialAction	http://www.cs.umd.edu/hcil/socialaction/	SocialAction is a social network analysis tool that integrates visualization and statistics to improve the analytical process.
200	SRI Law	http://www.ai.sri.com/~law/index.html	SRI Law is a Web-accessible tool where analysts and machines collaboratively perform link analysis by defining hierarchical and temporal patterns.
201	Stanford Entity Resolution Framework (SERF)	http://infolab.stanford.edu/serf/	The goal of the SERF project is to develop a generic infrastructure for Entity Resolution.
202	Stanford Named Entity Recognition (NER)	http://nlp.stanford.edu/ner/index.shtml	The Stanford Named Entity Recognizer (NER) is a Java implementation of a Conditional Random Field sequence model, together with well-engineered features for Named Entity Recognition.
203	Suggested Upper Merged Ontology (SUMO)	http://www.ontologyportal.org/	The Suggested Upper Merged Ontology (SUMO) and its domain ontologies form the largest formal public ontology in existence today.
204	SWOOP	http://code.google.com/p/swoop/	SWOOP is a tool for creating, editing, and debugging OWL ontologies. It was produced by the MIND lab at University of Maryland, College Park, but is now an open source project with contributors from all over.
205	Temis Insight Discoverer Extractor	http://www.temis.com/index.php?id=59&selt=1	Insight Discoverer Extractor is an information extraction server dedicated to the analysis of text document.
206	Temis Luxid	http://www.temis.com/index.php?id=70&selt=16	Luxid is a scalable solution giving immediate access to non obvious information and delivering industry-specific knowledge from internal and external data

ID	Product	URL	Description
			sources.
207	Teragram Entity Extraction	http://www.teragram.com/solutions/concepts_extr.htm	Multi-lingual natural language processing technologies that use the meaning of text to distill relevant information from vast amounts of data.
208	Termextractor (Beta)	http://lcl2.uniroma1.it/termextractor/demo.jsp	TermExtractor is a FREE software package for Terminology Extraction. The software helps a web community to extract and validate relevant domain terms in their interest domain, by submitting an archive of domain-related documents in any format."
209	Text Mining for Clementine	http://www.spss.com/text_mining_for_clementine/	Text Mining for Clementine is a text mining workbench that enables you to extract key concepts, sentiments, and relationships from textual or "unstructured" data and convert them to a structured format that can be used to create predictive models."
210	Text2Onto	http://ontoware.org/projects/text2onto/	Text2Onto is a framework for ontology learning from text.
211	Thetus Ontology Editor	http://www.thetus.com/products/detail.html	Thetus Ontology Editor provides an intuitive interface for examining and editing semantic categories and properties.
212	Thetus Publisher	http://www.thetus.com/products/	Thetus Publisher enables flexible and efficient modeling, discovery, sharing and re-use of data, metadata, and knowledge across sources, applications, disciplines and objectives.
213	ThinkMap SDK	http://www.thinkmap.com/thinkmapsdk.jsp;jsessionid=154B839561E3D3441724D4BB8B8E52BD	The Thinkmap SDK enables organizations to incorporate data-driven visualization technology into their enterprise Web application.
214	Tiburon Link Explorer	http://www.tiburoninc.com/solutions/link-analysis.asp	Tiburon LinkEXPLORER is a powerful analysis and visualization software solution that enables you to quickly uncover connections and associations critical to your investigation that you may have otherwise missed.
215	TIES (Trainable Information Extraction System)	http://tcc.itc.it/research/textec/tools-resources/ties.html	TIES automatically markups the documents with a predefined set of XML tags, exploiting markup rules automatically learned from a corpus previously annotated.
216	TopBraid Composer	http://www.topbraidcomposer.com/	TopBraid Composer is an enterprise-class platform for developing Semantic Web ontologies and building semantic applications.
217	TouchGraph Commercial	http://www.touchgraph.com/technology.html	TouchGraph is a set of interfaces for Graph Visualization using spring-layout and focus context techniques.
218	TouchGraph Open Source	http://sourceforge.net/projects/touchgraph	TouchGraph is a set of interfaces for Graph Visualization using spring-layout and focus context techniques.
219	T-Rex (Trainable Relation Extraction Framework)	http://www.aktors.org/technologies/trex/index.html	The Trainable Relation Extraction framework has been developed as a testbed for experimenting with several algorithms for relation extraction.

ID	Product	URL	Description
220	UCINET	http://www.analytictech.com/ucinet/ucinet.htm	UCINET is a comprehensive program for the analysis of social networks and other proximity data. The program contains dozens of network analytic routines.
221	UIMA	http://incubator.apache.org/uima/	UIMA is a framework and SDK for developing such applications. An example a UIMA application might ingest plain text and identify entities, such as persons, places, organizations; or relations, such as works-for or located-at.
222	Vertica Database Appliance	http://www.vertica.com/product/relational-database-management-system-overview	The Vertica Analytic Database lets you do for the business what you never thought was possible due to the performance limitations and high costs of traditional databases and proprietary analytic appliance hardware.
223	Visone	http://visone.info/about.php	Visone allows the user to load a graphical representation of a network and manipulate the links and the nodes.
224	Visual Browser	http://nlp.fi.muni.cz/projekty/vizualni_lexikon/	Visual Browser is a Java application that can visualize the data in RDF scheme.
225	VisualLinks	http://www.visualanalytics.com/products/visuaLinks/index.cfm	VisuaLinks is a platform-independent, graphical analysis tool used to discover patterns, trends, associations and hidden networks in any number and type of data sources.
226	VisualText	http://www.textanalysis.com/index.html	VisualText is an integrated development environment for building information extraction systems, natural language processing systems, and text analyzers.
227	VisuaLyzer	http://www.cnet.com.au/downloads/0,239030384,10437296s,00.htm	VisuaLyzer is an interactive tool for entering, visualizing and analyzing network data.
228	Wareman Software	http://www.woti.com/contact.cfm	White Oak Technologies, Inc. (WOTI) provides the next generation of solutions to massive, information-intensive, strategic intelligence challenges. WOTI's industry leading WAREMAN software has entity-resolved one of the worlds [sic] largest databases.
229	WebOnto	http://kmi.open.ac.uk/projects/webonto/	WebOnto is a Java applet coupled with a customized web server which allows users to browse and edit knowledge models over the web.
230	Xanalysis Indexer	http://www.intelligencesolutions.com.au/indexer.php	XANALYS Indexer automatically extracts relevant information from unstructured text, including entities, such as a person, company or an event, attributes, such as occupation, sex or company title and, relationships, such as located-at, works-for, involved.
231	YARS (Yet Another RDF Store).	http://sw.deri.org/2004/06/yars/	YARS (Yet Another RDF Store) is a data store for RDF in Java and allows for querying RDF.
232	YooName	http://www.yooname.com/Intro.html	YooName is Named Entity Recognition software based on semi-supervised learning. It identifies nine named entity categories that are split into more than 100 sub-categories.

By Category

Entity Extraction

ID Entity Extraction	Open Source	Commercial	Product	Company	Entity Extraction	Relationship Extraction	Semantic Integration	Entity Disambiguation	Knowledge Base	Visualization	Query	Analysis	Ontology/Data Model	Reference Data
2		√	21st Century Technologies Lynxeon	21st Century Technologies, Inc.	√	√		√						
5		√	AeroText Core Knowledge Base	Lockheed Martin	√	√								
9	√		ANNIE	The University of Sheffield	√									
11		√	Arabic Named Entity Extractor (ANEE)	Coltec (Computer and Language Technology)	√									
13		√	Attensity Extraction Engine	Attensity	√	√	√							
17	√		Balie (See also YooName)	University of Ottawa	√									
23		√	BBN Identifinder	BBN	√									
26		√	Bobcat	Decisive Analytics	√	√				√	√	√		
28		√	BullDoc	Trifeed	√									
29		√	BusinessObjects Text Analysis	BusinessObjects	√	√								
30		√	Carabao DeepAnalyzer	Digital Sonata	√									
31		√	Carabao Standard Free Edition	Digital Sonata	√									
35		√	Cicero	LCC (Language Computer Corporation)	√	√								
36		√	CiceroLite	LCC (Language Computer Corporation)	√	√								
38		√	Clarabridge Content Mining Platform	Clarabridge	√									
39	√		Classifier4J	Classifier4J	√									

ID Entity Extraction	Open Source	Commercial	Product	Company	Entity Extraction	Relationship Extraction	Semantic Integration	Entity Disambiguation	Knowledge Base	Visualization	Query	Analysis	Ontology/Data Model	Reference Data
41		√	ClearForest Extraction Modules	ClearForest	√	√								
42		√	COGITO Discover	Expert System	√									
46		√	Connexor Machinese Metadata	Connexor	√	√								
51		√	Cymfony Content Analysis (Info Extact Engine) Engine	TNS Media Intelligence/Cymfony	√	√								
57		√	DIANE Knowledge Services	Precipia	√					√	√	√		
58		√	Digital Reasoning GeoLocator	Digital Reasoning	√									
60	√		Dome	The University of Southampton	√									
61	√		ELIE	aidaf.net	√									
65		√	FASTUS	SRI	√									
68	√		FreeLing	FreeLing	√									
69	√		General Architecture for Text Engineering (GATE)	The University of Sheffield	√									
81		√	IBM Global Name Recognition	IBM	√									
89		√	Infogistics Xtractor	Infogistics	√	√								
93		√	Insightful InFact (Evri Solutions)	Evri Solutions	√									
97		√	Intelligenxia uReveal	Intelligenxia	√	√						√		
98		√	Interwoven MetaTagger	Interwoven	√									
100		√	Inxight Metadata Management System	Inxight	√	√								

ID Entity Extraction	Open Source	Commercial	Product	Company	Entity Extraction	Relationship Extraction	Semantic Integration	Entity Disambiguation	Knowledge Base	Visualization	Query	Analysis	Ontology/Data Model	Reference Data
103		√	Inxight SmartDiscovery Extraction Server (aka Analysis Server)	Inxight	√	√								
105		√	Inxight ThingFinder SDK	Inxight	√	√								
107		√	Janya Semantex	Janya	√	√								
109	√		jInFil	Claudio Giuliano	√									
111		√	Kaidara Text2data	Kaidara Software	√									
112		√	Kofax Capture	Kofax	√									
114		√	Lexalytics Salience Engine	Lexalytics	√	√								
119		√	LingPipe (aka ThreatTrackers)	Alias-I	√	√								
120		√	Linguamatics I2E	Linguamatics	√	√								
123		√	MarkLogic Server	MarkLogic	√				√		√			
124		√	Megaputer PolyAnalyst	Megaputer	√									
126		√	Melingo	Merlingo	√									
128		√	MetaCarta Geographic Text Search (GTS)	MetaCarta	√					√				
129		√	MetaCarta GeoTagger	MetaCarta	√					√				
132	√		Minor Third	William W. Cohen\Carnegie Mellon University	√					√				
140		√	NameFinder	AppTek	√									
143		√	NetOwl Extractor	SRA	√	√								
147	√		NMARKUP fact-file	University of Aberdeen	√									
159	√		OpenNLP	OpenNLP	√									
178		√	Readware Information	Readware	√									

ID Entity Extraction	Open Source	Commercial	Product	Company	Entity Extraction	Relationship Extraction	Semantic Integration	Entity Disambiguation	Knowledge Base	Visualization	Query	Analysis	Ontology/Data Model	Reference Data
			Processor											
183		√	Rosette Entity Extractor	Basis Technology	√									
190		√	SAS Text Miner	SAS	√	√								
195		√	Siderean Seamark MAPP	Siderean	√									
202	√		Stanford Named Entity Recognition (NER)	Stanford University	√									
205		√	Temis Insight Discoverer Extractor	Temis	√									
207		√	Teragram Entity Extraction	Teragram	√									
209		√	Text Mining for Clementine	SPSS	√					√				
215	√		TIES (Trainable Information Extraction System)	Claudio Giuliano	√									
221	√		UIMA	Apache (and IBM)	√									
226		√	VisualText	Text Analysis International	√									
230		√	Xanalysis Indexer	Xanalysis	√	√								
232	√		YooName	David Nadeau	√									
	15	**50**			√									

Relationship Extraction

ID Relationship Extraction	Open Source	Commercial	Product	Company	Entity Extraction	Relationship Extraction	Semantic Integration	Entity Disambiguation	Knowledge Base	Visualization	Query	Analysis	Ontology/Data Model	Reference Data
2		√	21st Century Technologies Lynxeon	21st Century Tecnologies, Inc.	√	√		√						
5		√	AeroText Core Knowledge Base	Lockheed Martin	√	√								
13		√	Attensity Extraction Engine	Attensity	√	√	√							
26		√	Bobcat	Decisive Analytics	√	√				√	√	√		
29		√	BusinessObjects Text Analysis	BusinessObjects	√	√								
35		√	Cicero	LCC (Language Computer Corporation)	√	√								
36		√	CiceroLite	LCC (Language Computer Corporation)	√	√								
41		√	ClearForest Extraction Modules	ClearForest	√	√								
46		√	Connexor Machinese Metadata	Connexor	√	√								
47		√	Content Analyst Latent Semantic Indexing	Content Analyst		√					√			
48	√		CORDER (COmmunity Relation Discovery by named Entity Recognition)	Jianhan Zhu (also Espotter)		√								
51		√	Cymfony Content Analysis (Info Extact Engine) Engine	TNS Media Intelligence/Cymfony	√	√								
89		√	Infogistics Xtractor	Infogistics	√	√								
97		√	Intelligenxia	Intelligenxia	√	√						√		

ID Relationship Extraction	Open Source	Commercial	Product	Company	Entity Extraction	Relationship Extraction	Semantic Integration	Entity Disambiguation	Knowledge Base	Visualization	Query	Analysis	Ontology/Data Model	Reference Data
			uReveal											
100		√	Inxight Metadata Management System	Inxight	√	√								
103		√	Inxight SmartDiscovery Extraction Server (aka Analysis Server)	Inxight	√	√								
105		√	Inxight ThingFinder SDK	Inxight	√	√								
107		√	Janya Semantex	Janya	√	√								
114		√	Lexalytics Salience Engine	Lexalytics	√	√								
117		√	Lexiquest Mine	SPSS		√				√				
119		√	LingPipe (aka ThreatTrackers)	Alias-I	√	√								
120		√	Linguamatics I2E	Linguamatics	√	√								
143		√	NetOwl Extractor	SRA	√	√								
190		√	SAS Text Miner	SAS	√	√								
219	√		T-Rex (Trainable Relation Extraction Framework)	The University of Sheffield		√								
230		√	Xanalysis Indexer	Xanalysis	√	√								
	2	**24**				√								

Semantic Integration

ID Semantic Integration	Open Source	Commercial	Product	Company	Entity Extraction	Relationship Extraction	Semantic Integration	Entity Disambiguation	Knowledge Base	Visualization	Query	Analysis	Ontology/Data Model	Reference Data
13		√	Attensity Extraction Engine	Attensity	√	√	√							
15		√	Attensity Solution Processors	Attensity			√							
21		√	BBN Asio Scout	BBN			√							
24		√	BBN Semantic Bridge for Relational Databases	BBN			√							
25		√	BBN Semantic Bridge for Web Services	BBN			√							
43		√	COGITO Intelligence	Expert System			√							
56		√	DIANE Core Server	Precipia			√							
66		√	Fetch Agent Platform	Fetch			√							
75	√		HP Labs Jena	HP Labs			√							
110	√		Joseki	HP Labs (see Jena)			√							
131		√	Metatomix Semantic Platform	Metatomix			√		√		√			
138		√	Modus Operandi Wave	Modus Operandi			√							
150		√	Ontoprise OntoStudio	Ontoprise			√						√	
152	√		Ontotext KIM Platform	Ontotext Semantic Technology Lab			√	√						
153	√		Ontotext ORDI SG	Ontotext Semantic Technology			√							

ID Semantic Integration	Open Source	Commercial	Product	Company	Entity Extraction	Relationship Extraction	Semantic Integration	Entity Disambiguation	Knowledge Base	Visualization	Query	Analysis	Ontology/Data Model	Reference Data
				Lab										
157	√		OpenLink Virtuoso Open Source	OpenLink Virtuoso Open Source			√							
158		√	OpenLink Virtuoso Universal Server	OpenLink Software			√	√	√		√			
179	√		RelationalOWL	Sourceforge Community			√							
181		√	Revelytix MatchIt	Revelytix			√							
192		√	Semantic Research Semantica SE	Semantica			√		√	√	√	√		
196		√	Siderean Seamark Navigator	Siderean			√		√		√	√		
	6	**15**					√							

Entity Disambiguation

ID Entity Disambiguation	Open Source	Commercial	Product	Company	Entity Extraction	Relationship Extraction	Semantic Integration	Entity Disambiguation	Knowledge Base	Visualization	Query	Analysis	Ontology/Data Model	Reference Data
2		√	21st Century Technologies Lynxeon	21st Century Tecnologies, Inc.	√	√		√						
90		√	Initiate Customer	Initiate Systems				√	√	√	√	√		
91		√	Initiate Master Data Service	Initiate Systems				√	√	√	√	√		
92		√	Initiate Organization	Initiate Systems				√	√	√	√	√		
152	√		Ontotext KIM Platform	Ontotext Semantic Technology Lab			√	√						
158		√	OpenLink Virtuoso Universal Server	OpenLink Software			√	√	√		√			
184		√	Rosette Name Indexer	Basis Technology				√	√		√			
185		√	SaffronScope	Saffron Technology Inc.				√			√	√		
186		√	SaffronWeb	Saffron Technology Inc.				√			√	√		
201	√		Stanford Entity Resolution Framework (SERF)	Stanford University				√						
228		√	Wareman Software	White Oak Technologies				√						
	2	**9**						√						

Knowledge Base

ID Knowledge Base	Open Source	Commercial	Product	Company	Entity Extraction	Relationship Extraction	Semantic Integration	Entity Disambiguation	Knowledge Base	Visualization	Query	Analysis	Ontology/Data Model	Reference Data
4	√		3store	The University of Southampton					√		√			
6		√	AeroText Knowledge Base Engine	Lockheed Martin					√		√			
7		√	AllegroGraph	Franz, Inc.					√		√			
20		√	BBN Asio Parliament	BBN					√		√			
27	√		Brahms	The University of Georgia					√		√			
49		√	Cyc Knowledge Base	Cycorp					√		√		√	
50	√		Cycorp OpenCyc	Cycorp					√		√		√	
53	√		D2R Server	Freie Universitat Berlin					√		√			
76	√		HP Labs SDB	HP Labs (see Jena)					√		√			
85	√		IBM Semantic Layered Research Program (Boca)	IBM					√		√			
90		√	Initiate Customer	Initiate Systems				√	√	√	√	√		
91		√	Initiate Master Data Service	Initiate Systems				√	√	√	√	√		
92		√	Initiate Organization	Initiate Systems				√	√	√	√	√		
96		√	Intellidimension RDF Gateway	IntelliDimensions					√		√			
108	√		Jena with PostgreSQL	Postgresql Community					√		√			
113	√		Kowari	Kowari Community					√		√			
123		√	MarkLogic Server	MarkLogic	√				√		√			
131		√	Metatomix	Metatomix			√		√		√			

ID Knowledge Base	Open Source	Commercial	Product	Company	Entity Extraction	Relationship Extraction	Semantic Integration	Entity Disambiguation	Knowledge Base	Visualization	Query	Analysis	Ontology/Data Model	Reference Data
			Semantic Platform											
139	√		Mulgara (see Kowari)	Mulgra.org					√		√			
151	√		Ontotext BigOWLIM (see also OWLIM) Semantic Repository	Ontotext Semantic Technology Lab					√		√			
154	√		Ontotext OWLIM (see also BigOWLIM)	Ontotext Semantic Technology Lab					√		√			
156	√		Open Anzo	Open Anzo					√		√			
158		√	OpenLink Virtuoso Universal Server	OpenLink Software			√	√	√		√			
160		√	Oracle 11g (Spatial)	Oracle					√		√			
175	√		RDF2Go	FZI, Karlsruhe, Germany					√		√			
177	√		RDFStore	RDFStore					√		√			
184		√	Rosette Name Indexer	Basis Technology				√	√		√			
192		√	Semantic Research Semantica SE	Semantica			√		√	√	√	√		
193	√		Semantic Web RDF Library for C#/.NET	Joshua Tauberer					√		√			
194	√		Sesame	Aduna					√		√			
196		√	Siderean Seamark Navigator	Siderean			√		√		√	√		
212		√	Thetus Publisher	Thetus					√	√	√	√		
222		√	Vertica Database Appliance	Vertica					√		√			
231	√		YARS (Yet Another RDF Store).	Deri.org					√		√			

ID Knowledge Base	Open Source	Commercial	Product	Company	Entity Extraction	Relationship Extraction	Semantic Integration	Entity Disambiguation	Knowledge Base	Visualization	Query	Analysis	Ontology/Data Model	Reference Data
	17	**17**							√					

Visualization

ID Visualization	Open Source	Commercial	Product	Company	Entity Extraction	Relationship Extraction	Semantic Integration	Entity Disambiguation	Knowledge Base	Visualization	Query	Analysis	Ontology/Data Model	Reference Data
3		√	21st Century Technologies Threat Detection and Analysis (TMODS)	21st Century Tecnologies, Inc.						√		√		
10	√		Apache Agora	Stefano Mazzocchi						√				
12		√	Attensity Explore\Analytics	Attensity						√		√		
16		√	AXIS	Overwatch Textron Systems						√				
26		√	Bobcat	Decisive Analytics	√	√				√	√	√		
32		√	Centrifuge	Tildenwoods						√		√		
57		√	DIANE Knowledge Services	Precipia	√					√	√	√		
62		√	Endeca Information Access Platform	Endeca						√	√	√		
67		√	FMS Sentinel Visualizer	FMS Advanced Systems Group						√		√		
70	√		Graphl	Grapl Open Source						√				
71	√		Graphviz	Graphviz						√				
73	√		Guess, The Graph Exploration System	Graphexploration						√				
77	√		HyperTree Java Library	HyperTree Java Library						√				
78		√	i2 Analyst's Notebook	i2 Inc.						√		√		
79		√	i2 ChartExplorer	i2 Inc.						√				
88		√	Inflow	Orgnet (Valdis Krebs)						√		√		
90		√	Initiate Customer	Initiate Systems				√	√	√	√	√		
91		√	Initiate Master	Initiate Systems				√	√	√	√	√		

ID Visualization	Open Source	Commercial	Product	Company	Entity Extraction	Relationship Extraction	Semantic Integration	Entity Disambiguation	Knowledge Base	Visualization	Query	Analysis	Ontology/Data Model	Reference Data
			Data Service											
92		√	Initiate Organization	Initiate Systems				√	√	√	√	√		
94		√	Insightful Miner	Insightful Corp.						√		√		
104		√	Inxight SmartDiscovery VizServer	Inxight						√	√	√		
106	√		IsaViz: A Visual Authoring Tool for RDF	Emmanuel Pietriga						√				
117		√	Lexiquest Mine	SPSS		√				√				
118	√		LibSea	Caida.org						√				
122		√	Lucid Threat Management System	Dulles Research						√		√		
128		√	MetaCarta Geographic Text Search (GTS)	MetaCarta	√					√				
129		√	MetaCarta GeoTagger	MetaCarta	√					√				
132	√		Minor Third	William W. Cohen\Carnegie Mellon University	√					√				
142		√	NetMiner	Cyram						√				
144		√	NetOwl InstaLink	SRA						√		√		
145		√	NetOwl TextMiner	SRA						√	√	√		
161	√		Organizational Risk Assessment (ORA)	Carnegie Mellon University						√		√		
164	√		Pajek	University of Ljubljana, Slovenia						√		√		
165		√	Paladin	Metron						√		√		
166		√	Palantir	Palantir Technologies						√		√		
167	√		Picoolo Toolkit	University of Maryland						√				
169	√		Prefuse (see also SocialAction)	Prefuse.org						√				

ID Visualization	Open Source	Commercial	Product	Company	Entity Extraction	Relationship Extraction	Semantic Integration	Entity Disambiguation	Knowledge Base	Visualization	Query	Analysis	Ontology/Data Model	Reference Data
171	√		Proximity	University of Mass\Amherst						√		√		
173	√		RapidMiner (formerly YALE)	Rapid-i						√		√		
174	√		RDF Gravity (RDF Graph Visualization Tool)	Salzburg Research						√	√			
192		√	Semantic Research Semantica SE	Semantica			√		√	√	√	√		
199	√		SocialAction	University of Maryland						√				
200		√	SRI Law	SRI						√		√		
209		√	Text Mining for Clementine	SPSS	√					√				
212		√	Thetus Publisher	Thetus					√	√	√	√		
213		√	ThinkMap SDK	Thinkmap, Inc.						√		√		
214		√	Tiburon Link Explorer	Tuburon						√		√		
217		√	TouchGraph Commercial	TouchGraph Commercial						√				
218	√		TouchGraph Open Source	TouchGraph Open Source						√				
220		√	UCINET	Analytictech						√				
223		√	Visone	University of Konstanz and Karlsruhe						√		√		
224		√	Visual Browser	Visual Browser						√				
225		√	VisualLinks	VisualAnalytics						√				
227	√		VisuaLyzer	Medical Decision Logic						√				
	18	**36**								√				

Query

ID Query	Open Source	Commercial	Product	Company	Entity Extraction	Relationship Extraction	Semantic Integration	Entity Disambiguation	Knowledge Base	Visualization	Query	Analysis	Ontology/Data Model	Reference Data
1		√	21st Century Technologies Large Scale Data Searching	21st Century Tecnologies, Inc.							√			
4	√		3store	The University of Southampton					√		√			
6		√	AeroText Knowledge Base Engine	Lockheed Martin					√		√			
7		√	AllegroGraph	Franz, Inc.					√		√			
14		√	Attensity Search	Attensity							√			
20		√	BBN Asio Parliament	BBN					√		√			
22		√	BBN Asio Semantic Query Decomposition	BBN							√			
26		√	Bobcat	Decisive Analytics	√	√				√	√	√		
27	√		Brahms	The University of Georgia					√		√			
37		√	Clarabridge Business Intelligence Search	Clarabridge							√	√		
44		√	COGITO Semantic Search	Expert System							√			
47		√	Content Analyst Latent Semantic Indexing	Content Analyst		√					√			
49		√	Cyc Knowledge Base	Cycorp					√		√		√	
50	√		Cycorp OpenCyc	Cycorp					√		√		√	
53	√		D2R Server	Freie Universitat Berlin					√		√			

ID Query	Open Source	Commercial	Product	Company	Entity Extraction	Relationship Extraction	Semantic Integration	Entity Disambiguation	Knowledge Base	Visualization	Query	Analysis	Ontology/Data Model	Reference Data
57		√	DIANE Knowledge Services	Precipia	√					√	√	√		
62		√	Endeca Information Access Platform	Endeca						√	√	√		
76	√		HP Labs SDB	HP Labs (see Jena)					√		√			
85	√		IBM Semantic Layered Research Program (Boca)	IBM					√		√			
90		√	Initiate Customer	Initiate Systems				√	√	√	√	√		
91		√	Initiate Master Data Service	Initiate Systems				√	√	√	√	√		
92		√	Initiate Organization	Initiate Systems				√	√	√	√	√		
96		√	Intellidimension RDF Gateway	IntelliDimensions					√		√			
99		√	Interwoven Universal Search	Interwoven							√			
101		√	Inxight Search Extender for Google Desktop	Inxight							√			
102		√	Inxight SmartDiscovery Awareness Server	Inxight							√	√		
104		√	Inxight SmartDiscovery VizServer	Inxight						√	√	√		
108	√		Jena with PostgreSQL	Postgresql Community					√		√			
113	√		Kowari	Kowari Community					√		√			
123		√	MarkLogic Server	MarkLogic	√				√		√			
131		√	Metatomix Semantic Platform	Metatomix			√		√		√			

ID Query	Open Source	Commercial	Product	Company	Entity Extraction	Relationship Extraction	Semantic Integration	Entity Disambiguation	Knowledge Base	Visualization	Query	Analysis	Ontology/Data Model	Reference Data
139	√		Mulgara (see Kowari)	Mulgra.org					√		√			
145		√	NetOwl TextMiner	SRA						√	√	√		
151	√		Ontotext BigOWLIM (see also OWLIM) Semantic Repository	Ontotext Semantic Technology Lab					√		√			
154	√		Ontotext OWLIM (see also BigOWLIM)	Ontotext Semantic Technology Lab					√		√			
156	√		Open Anzo	Open Anzo					√		√			
158		√	OpenLink Virtuoso Universal Server	OpenLink Software			√	√	√		√			
160		√	Oracle 11g (Spatial)	Oracle					√		√			
174	√		RDF Gravity (RDF Graph Visualization Tool)	Salzburg Research						√	√			
175	√		RDF2Go	FZI, Karlsruhe, Germany					√		√			
177	√		RDFStore	RDFStore					√		√			
182		√	Rosette Cross-Language Toolkit	Basis Technology							√			
184		√	Rosette Name Indexer	Basis Technology				√	√		√			
185		√	SaffronScope	Saffron Technology Inc.				√			√	√		
186		√	SaffronWeb	Saffron Technology Inc.				√			√	√		
188		√	SAS Enterprise Miner	SAS							√	√		
192		√	Semantic Research Semantica SE	Semantica			√		√	√	√	√		
193	√		Semantic Web RDF Library for C#/.NET	Joshua Tauberer					√		√			

ID Query	Open Source	Commercial	Product	Company	Entity Extraction	Relationship Extraction	Semantic Integration	Entity Disambiguation	Knowledge Base	Visualization	Query	Analysis	Ontology/Data Model	Reference Data
194	√		Sesame	Aduna					√		√			
196		√	Siderean Seamark Navigator	Siderean			√		√		√	√		
206		√	Temis Luxid	Temis							√	√		
212		√	Thetus Publisher	Thetus					√	√	√	√		
222		√	Vertica Database Appliance	Vertica					√		√			
231	√		YARS (Yet Another RDF Store).	Deri.org					√		√			
	18	**36**									√			

Analysis

ID Analysis	Open Source	Commercial	Product	Company	Entity Extraction	Relationship Extraction	Semantic Integration	Entity Disambiguation	Knowledge Base	Visualization	Query	Analysis	Ontology/Data Model	Reference Data
3		√	21st Century Technologies Threat Detection and Analysis (TMODS)	21st Century Tecnologies, Inc.						√		√		
12		√	Attensity Explore\Analytics	Attensity						√		√		
26		√	Bobcat	Decisive Analytics	√	√				√	√	√		
32		√	Centrifuge	Tildenwoods						√		√		
37		√	Clarabridge Business Intelligence Search	Clarabridge							√	√		
40		√	ClearForest Analytics	ClearForest								√		
52		√	Cymfony Orchestra	TNS Media Intelligence/Cymfony								√		
57		√	DIANE Knowledge Services	Precipia	√					√	√	√		
59		√	Digital Reasoning Interceptor	Digital Reasoning								√		
62		√	Endeca Information Access Platform	Endeca						√	√	√		
67		√	FMS Sentinel Visualizer	FMS Advanced Systems Group						√		√		
78		√	i2 Analyst's Notebook	i2 Inc.						√		√		
88		√	Inflow	Orgnet (Valdis Krebs)						√		√		
90		√	Initiate Customer	Initiate Systems				√	√	√	√	√		
91		√	Initiate Master Data Service	Initiate Systems				√	√	√	√	√		
92		√	Initiate Organization	Initiate Systems				√	√	√	√	√		
94		√	Insightful Miner	Insightful Corp.						√		√		

ID Analysis	Open Source	Commercial	Product	Company	Entity Extraction	Relationship Extraction	Semantic Integration	Entity Disambiguation	Knowledge Base	Visualization	Query	Analysis	Ontology/Data Model	Reference Data
97		√	Intelligenxia uReveal	Intelligenxia	√	√						√		
102		√	Inxight SmartDiscovery Awareness Server	Inxight							√	√		
104		√	Inxight SmartDiscovery VizServer	Inxight						√	√	√		
115		√	Leximancer Professional	Leximancer								√		
116		√	Leximancer Server	Leximancer								√		
122		√	Lucid Threat Management System	Dulles Research						√		√		
125		√	Megaputer TextAnalyst	Megaputer								√		
144		√	NetOwl InstaLink	SRA						√		√		
145		√	NetOwl TextMiner	SRA						√	√	√		
161	√		Organizational Risk Assessment (ORA)	Carnegie Mellon University						√		√		
164	√		Pajek	University of Ljubljana, Slovenia						√		√		
165		√	Paladin	Metron						√		√		
166		√	Palantir	Palantir Technologies						√		√		
171	√		Proximity	University of Mass\Amherst						√		√		
173	√		RapidMiner (formerly YALE)	Rapid-i						√		√		
185		√	SaffronScope	Saffron Technology Inc.				√			√	√		
186		√	SaffronWeb	Saffron Technology Inc.				√			√	√		
188		√	SAS Enterprise Miner	SAS							√	√		
189		√	SAS Model Manager	SAS								√		

ID Analysis	Open Source	Commercial	Product	Company	Entity Extraction	Relationship Extraction	Semantic Integration	Entity Disambiguation	Knowledge Base	Visualization	Query	Analysis	Ontology/Data Model	Reference Data
192		√	Semantic Research Semantica SE	Semantica			√		√	√	√	√		
196		√	Siderean Seamark Navigator	Siderean			√		√		√	√		
200		√	SRI Law	SRI						√		√		
206		√	Temis Luxid	Temis							√	√		
212		√	Thetus Publisher	Thetus					√	√	√	√		
213		√	ThinkMap SDK	Thinkmap, Inc.						√		√		
214		√	Tiburon Link Explorer	Tuburon						√		√		
223		√	Visone	University of Konstanz and Karlsruhe						√		√		
	4	**40**										√		

Ontology/Data Model

ID Ontology\Data Model	Open Source	Commercial	Product	Company	Entity Extraction	Relationship Extraction	Semantic Integration	Entity Disambiguation	Knowledge Base	Visualization	Query	Analysis	Ontology/Data Model	Reference Data
8		√	Altova Semantic Web Tool	Altova									√	
18	√		Basic Formal Ontology (BFO)	Barry Smith and Pierre Grenon									√	
19		√	BBN Asio Cartographer	BBN									√	
33		√	Ceryph Insight	Ceryph									√	
45	√		Common Terrorism Information Sharing Standards (CTISS)	Information Sharing Environment (ISE)									√	
49		√	Cyc Knowledge Base	Cycorp					√		√		√	
50	√		Cycorp OpenCyc	Cycorp					√		√		√	
54	√		DERI Ontology Management Environment (DOME)	DOME Open Source									√	
55	√		Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE)	Nicola Guarino									√	
72	√		GrOwl	University of Vermont									√	
74	√		Hozo Ontology Editor	Osaka University									√	
83	√		IBM Integrated Ontology Development Toolkit (IODT)	IBM									√	
86	√		IBM Web Ontology Manager	IBM									√	

ID Ontology\Data Model	Open Source	Commercial	Product	Company	Entity Extraction	Relationship Extraction	Semantic Integration	Entity Disambiguation	Knowledge Base	Visualization	Query	Analysis	Ontology/Data Model	Reference Data
87	√		IHMC CmapTools	IHMC									√	
95		√	Intellidimension InferEd	IntelliDimensions									√	
121		√	LinKFFactory	Language & Computing									√	
127		√	Meta Integration Model Bridge (MIMB) "Metadata Integration" Solution	MetaIntegration									√	
130	√		Metatomix m3t4.studio Semantic Tookkit	Metatomix Open Source									√	
137	√		Model Futures OWL Editor	Model Futures									√	
148	√		oBrowse	oBrowse									√	
149		√	Ontology Works Integrated Ontology Development Environment	Ontology Works, Inc.									√	
150		√	Ontoprise OntoStudio	Ontoprise			√						√	
155	√		OntoWare Oyster	OntoWare									√	
162	√		OWL Verbalizer	OWL Verbalizer									√	
163	√		OwlSight	Clark & Parsia									√	
168	√		pOwl	pOwl									√	
170	√		Protégé	Stanford University									√	
172	√		RAP NetAPI	Phil Dawes									√	
176	√		RDFe	RDFe									√	
180		√	Revelytix Knoodl	Revelytix									√	
187		√	Sandpiper Visual Ontology	Sandpiper Software, Inc.									√	

ID Ontology\Data Model	Open Source	Commercial	Product	Company	Entity Extraction	Relationship Extraction	Semantic Integration	Entity Disambiguation	Knowledge Base	Visualization	Query	Analysis	Ontology/Data Model	Reference Data
			Modeler											
191		√	SchemaLogic Enterprise Suite	SchemaLogic									√	
197		√	Smartlogic Ontology Manager	Smartlogic									√	
198	√		Snoogle	Snoogle Open Source									√	
203	√		Suggested Upper Merged Ontology (SUMO)	Adam Pease									√	
204	√		SWOOP	Mind Lab University of Maryland									√	
208	√		Termextractor (Beta)	Termextractor (Beta)									√	
210	√		Text2Onto	Ontoware Community									√	
211		√	Thetus Ontology Editor	Thetus									√	
216		√	TopBraid Composer	TopQuadrant									√	
229	√		WebOnto	John Domingue									√	
	26	**15**											√	

Reference Model

ID Reference Model	Open Source	Commercial	Product	Company	Entity Extraction	Relationship Extraction	Semantic Integration	Entity Disambiguation	Knowledge Base	Visualization	Query	Analysis	Ontology/Data Model	Reference Data
34	√		CIA World FactBook	CIA										√
64		√	Factiva Taxonomy Warehouse	Dow Jones Factiva										√
133	√		MIPT State Department list of Foreign Terrorist Organizations (FTO)	Memorial Institute for the Prevention of Terrorism (MIPT)										√
134	√		MIPT State Department list of selected Other Terrorist Organizations (OTO)	Memorial Institute for the Prevention of Terrorism (MIPT)										√
135	√		MIPT State Department Terrorist Exclusion List (TEL)	Memorial Institute for the Prevention of Terrorism (MIPT)										√
136	√		MIPT Terrorism Knowledge Base (TKB)	Memorial Institute for the Prevention of Terrorism (MIPT)										√
141	√		NCTC Worldwide Incidents Tracking System (WITS)	NCTC										√
146	√		NGA GEOnet Names Server (GNS)	National Geospatial Intellligence Agency (NGIC)										√
	7	**1**												√

(U//FOUO) Appendix D. Government Systems Descriptions (U)

(U) The following are the descriptions of the government systems that contributed to the conclusions of Section 5. Whereas these are probably not the only systems in the Intelligence Community that have functionality that matches up with that required for an eventual Catalyst system, we believe that they represent a reasonable cross-section of the kinds of systems being developed. This study did focus primarily on the national intelligence agencies, so we particularly do not believe that we have identified all the potential systems of interest in the Commands and other agencies not located in the greater Washington, DC metropolitan area. As stated in the conclusions, there may be other such systems that should be identified and described, and the conclusions of Section 5 should be reviewed with any additional data in mind. Also, over time additional capabilities and systems are likely to be developed, and again the conclusions might evolve based on these additions.

(U//FOUO) Program Name: Project AETHER

(U//FOUO) **Sponsoring organization:** Office of Naval Intelligence (ONI)'s Advanced Maritime Analysis Cell (AMAC), and the Intelligence Advanced Research Projects Activity (IARPA).

(U//FOUO) **Performing contractor(s):** SAIC with support from Booz, Allen and others.

(U//FOUO) **Gov't POC Phone Number & E-mail Address:** LCDR Jim Ford, ONI, (301) 669-2050, james.p.ford@ugov.gov.

(U//FOUO) **Contractor POC Phone Number & E-mail Address:** David Lippert, SAIC, (703) 276-3117, david.r.lippert@saic.com.

(U//FOUO) **Abstract description:** AETHER will enable analysts to use various information sources to correlate seemingly disparate entities and relationships, to identify networks of interest, and to detect patterns. The AETHER architecture will enable advanced analysis of billions of entities and relationships, providing analysts the capability to seamlessly manage their information sources and produce reliable hypothesis and intelligence reports.

(U//FOUO) List of primary Aether capabilities:

- Import
- csv files
- NGA gazetteer data
- CIA World factbook files
- Harvest
- Multi-file harvest from zip files
- Google harvest
- Pdf's, txt, rss
- Annotate
- Create, edit, & delete entities and relationships
- Search & Organize
- Text based search of documents
- Search across Problems of Interest (POIs) and global datasets
- Document organizer with full provenance
- Share datasets and POIs
- Jungle Browser
- View & export semantic graphs
- Expand and collapse data

AETHER is a follow-on of the Proteus program at ONI, which no longer exists.

(U//FOUO) **Intended users:** Intelligence analysts, currently in the maritime domain, but it can be adapted to any other intelligence analysis domain.

(U//FOUO) **Catalyst functionality included**:

Entity extraction	√
Relationship extraction	√ (manual only)
Metadata management	√
Semantic entity integration	√
Entity disambiguation	√
Entity knowledge base	√
Visualization	√
Query	√
Knowledge management	

(U//FOUO) **Sources of input data:** Results of Google searches, rss feeds, pdf & text documents, NGA gazetteer data, CIA world factbook data, csv files.

(U//FOUO) **Scale of current implementation:** The current RDF throughput is approx. 20k triplets per second and the development team is steadily improving the scalability. The current number of entities and relationships has increased from 1000 to 2500 and any performance lag is associated with the user interface and not with the knowledge base. The current ontology has been designed for simplicity over deepness for analysis, with about 9 classes (and few subclasses).

(U//FOUO) **Status of system:** In development for only about 6 months.

(U//FOUO) **Where deployed:** Initial early adopters of this capability will include ONI AMAC and TRIDENT, NCTC, USSOCOM, DIA JITF-CT and COMFIFTHFLT and possibly NSA.

(U//FOUO) **COTS/OS/GOTS used:** Open source components include Bigdata (RDF Store), Sesame 1.x (RDF Framework), Lucene (indexer), Mysql (content repository), Jung (foundation for graphical semantic visualization), JBOSS (foundation of web interface), and JMS (foundation of workflow manager for processing unstructured data). GOTS includes Aether's harvester, text extractor (entities only; relationships are extracted manually by the user while interacting with AETHER), annotator, and evidence viewer. It is being integrated with IARPA's BlackBook.

(U//FOUO) **Size of development effort:** Less than 10 people.

(U//FOUO) **User experiences:**

(U//FOUO) **Plans for continued development:** Expand the scalability of Aether and move toward an enterprise level and hardened system for operational usage. Software is moving to Sesame 2.x and SPARQL (in March 2008) and replacing MySQL with an internal database using Java.

(U//FOUO) **Lessons learned**:

(U//FOUO) Challenges

- Data perceptions and concerns
- General integration issues

- Contributors to Aether's success
- Limited automated entity extraction
- Unclassified development and integration
- Analyst driven requirements
- Strong team and active customer involvement

(U//FOUO) Program Name: APSTARS

(U//FOUO) **Sponsoring organization**: NSA/T1222

(U//FOUO) **Performing contractor(s)**: SAIC, i_SW, Chiron Technology Services, Smearman IT

(U//FOUO) **Gov't POC Phone Number & E-mail Address**: Brian Maddox, 240/373-8697, john.b.maddox@ugov.gov

(U//FOUO) **Contractor POC Phone Number & E-mail Address**: Brad Bebee, 571/265-5508, beebs@iswcorp.com

(U//FOUO) **Abstract description**: Semantic integration of data from multiple sources in support of intelligence processing. Some important sources are from the Internet, for which semi-structured web page scraping is done. Included is a One-Way Transfer capability to move this data up to the classified network. The Internet harvesting and OWT will keep the APSTARS name and will transition to an enterprise service, while the remaining capability to semantically integrate and analyze data will become a new program, called HERESYITCH.

(U//FOUO) **Intended users**: Multiple distinct mission areas within NSA.

(U//FOUO) **Catalyst functionality included**:

Entity extraction	√
Relationship extraction	√ (but from semi-structured web pages, not text)
Metadata management	
Semantic entity integration	√
Entity disambiguation	√
Entity knowledge base	√
Visualization	√
Query	√
Knowledge management	√

(U//FOUO) **Sources of input data**: Web pages and databases on Internet. Moving towards classified sources on NSANet.

(U//FOUO) **Scale of current implementation**: Testing has been done up to approx. 120M triples. There are ~100 beta users, but mainly of the harvesting and OWT.

(U//FOUO) **Status of system**: Expect parts to be operational within 1-2 months. ATOs for three components: harvester and storage/analysis at PL2, One Way Transfer at PL5.

(U//FOUO) **Where deployed**: NSANet.

(U//FOUO) **COTS/OS/GOTS used**: BrightPlanet for deep web harvesting, Kapow for web scraping, Siderean Seamark for triple store, analytics, and visualization. Heritrix open source crawling software from archive.org.

(U//FOUO) **Size of development effort**: ~5 FTEs.

(U//FOUO) **User experiences**: Positive for harvesting and OWT part, none yet for semantic integration part.

(U//FOUO) **Plans for continued development**: Moving to Oracle for persistent storage and Seamark for analytics. While generally pleased with Seamark, Siderean is moving to a service model for their products, while will not work well for the classified environment. Getting into unstructured data, for which will use capabilities from the Center for Content Extraction (CCE).

(U//FOUO) **Lessons learned**: The philosophy of using the fewest semantics to get simple functions done pays off. The lighter weight approach is more scalable and more accepting by users.

(U//FOUO) There is a core ontology that gets extended for each mission area. The ontology is at the RDFS plus some OWL axioms level of expressivity.

(U//FOUO) Program Name: BLACKBOOK2

(U//FOUO) **Sponsoring organization:** IARPA/ Knowledge Discovery and Dissemination (KDD) Program.

(U//FOUO) **Performing contractor(s):** Johns Hopkins University/Applied Physics Laboratory, SRA International, CACI.

(U//FOUO) **Gov't POC Phone Number & E-mail Address:** H. "Buster" Fields, Program Manager, 240-373-5309, hlfield@nsa.gov.

(U//FOUO) **Contractor POC Phone Number & E-mail Address:** Dr. John "Jack" Callahan, Senior Technical Advisor**,** 443-778-3674, john.callahan@jhuapl.edu**;** Emerson Brooks, Software Team Manager, 443-656-7312, emerson_brooks@sra.com.

(U//FOUO) **Abstract Description:** BLACKBOOK2 is a Semantic Web application that gives IC analysts the ability to 1) Upload one or more ontologies via Ontology Manager, 2) create and edit entities and their properties, via Entity Manager, 3) make statements about relationships between entities, and annotate these assertions with confidence values and security classifications, via Relationship Manager, 4) define workflow process definitions, via Workflow Manager, 5) share assertions and graph results with colleagues, via Workspace Manager, and 6) view relevant information using social network analysis, temporal, and geospatial techniques.

(U//FOUO) The BLACKBOOK2 infrastructure exposes core capabilities via web services. All assertions made by analysts are automatically marked with user name and agency affiliation, and stored in an internal knowledge base with date/time stamp. BLACKBOOK2 is a server-based thin-client that uses "best-of-breed" open-source technologies. Using PKI certs with corporate authentication services, BLACKBOOK2 is accredited for network security PL3+.

(U//FOUO) **Intended users:** IC Analysts (multi-INT), business intelligence users, and academic, commercial, and government researchers.

(U//FOUO) **Catalyst functionality included:**

Entity extraction	√
Relationship extraction	√
Metadata management	√
Semantic entity integration	√
Entity disambiguation	√
Entity knowledge base	√
Visualization	√
Query	√
Knowledge management	√

(U//FOUO) **Sources of input data:** BLACKBOOK2 connects to 8 data sources: 1) Anubis (bio-equipment and bio-scientists), 2) Monterey (terrorist incidents), 3) Sandia (terrorist profiles), 4) Medline (bio-data), 5) Artemis (bio-weapons proliferation), 6)

BACWORTH (biological and chemical weapons), 7) the 9/11 Commission Report, and 8) an NGA Google-maps service.

(U//FOUO) **Scale of current implementation:** The largest data source connected to BLACKBOOK2 is Medline, at 290 GBytes in size. Transformed to RDF and directly ingested, Medline represents 1.5 million entities and 2.5 billion individual statements. These statements are Lucene indexed using 7.7 billion indices and Jena indexed using 247 billion indices.

(U//FOUO) **Status of system:** Still under evaluation, with deployments to a number of agencies. BLACKBOOK2 source code is available for download from the Blackbook wiki on the Internet, and numerous developers in academic, commercial, and government settings are also contributing to BLACKBOOK2 development.

(U//FOUO) **Where deployed:** BLACKBOOK2 is currently deployed on JWICS and NSAnet, as well as RDEC. Instances of BLACKBOOK2 are also deployed on a number of test and evaluation network enclaves at CIA, NSA, NCTC, NGIC, and JIEDDO.

(U//FOUO) **COTS/OS/GOTS used:** Jena RDF Triple Store, Lucene Index Engine, D2RQ, JUNG network graph visualization, NetOwl entity extraction, I-2 Analyst Notebook with rLink plug-in (rLink enables Analyst Notebook to communicate with/display contents of RDEC's EDB database), Network Workbench from School of Library and Information Science, Indiana University.

(U//FOUO) **Size of development effort:** Approx. 12 FTEs for core development team.

(U//FOUO) **User experiences:** Too soon to report.

(U//FOUO) **Plans for continued development:** BLACKBOOK2 is still in the maturation phase. Current development is focused on enhancing integration points for data sources, algorithms, and visualization. Additionally, a peer-to-peer capability to allow secure remote invocation across multiple BLACKBOOK2 instances across network domains is under development. Future deployments are planned for A-SPACE-U and A-SPACE-R.

(U//FOUO) **Details:** BLACKBOOK2 consists of three integration points; 1) Data Sources, 2) Algorithms, and 3) Visualizations.

(U//FOUO) Data Source Integration. The process for integrating an *internal* data source requires that a copy of the data be imported into BLACKBOOK2. An external RDBMS data source is transformed into RDF, then directly stored and Lucene indexed. BLACKBOOK2 uses the RDF as its abstract data model and Jena as its RDF implementation. The current approach to integrating an *external* RDBMS data source uses D2RQ, an open-source bridging technology, which maps from an RDBMS schema to RDF. This allows BLACKBOOK2 to "see" any external database as and RDF store. There is a performance penalty for this mapping, however, this approach has the advantage of requiring no modifications to the original data, and avoids the issues of synchronization associated with re-hosting the data.

(U//FOUO) Algorithm Integration. BLACKBOOK2 extensibility is achieved through the concept of Algorithms. Algorithms provide the means for injecting new and value added functionality; like data filtering, transformation, and manipulation. A few examples functions are queries, dips, expands, and materialization.

(U//FOUO) When an Analyst performs a *keyword* query, BLACKBOOK2 searches all available resources and returns one or more Uniform Resource Identifiers (URI) that 'point to' RDF document(s). For example, a keyword query for "Smith," may return URIs to a person with the last name of "Smith" or a person who lives on a street named "Smith Street."

(U//FOUO) Subsequent to a keyword query, an Analyst can perform a *Dip* query – whereby name/value attributes for selected item(s) are matched across all data sources. For instance, a *person* entity might have a last name attribute with the value "Smith". BLACKBOOK2 will look in all of its data sources for last name attributes and when a match is found, will return results from that data source for *person* entities that have last names of "Smith".

(U//FOUO) The *Expand* function is a query performed within an entity's data set that returns all of the entities directly related to the entity. For instance, if a person entity is expanded, it would be reasonable to expect that the person's attributes such as age, sex, occupation, etc. would be part of the search results returned.

(U//FOUO) Lastly, the *Materialize* function is an adjunct capability to the first three mentioned above. As stated earlier, the results returned for the Keyword, Dip and Expand functions are URIs or 'pointers' to the data. Materialize returns the source data pointed to by these URIs. For example, a URI may point to a PDF document residing in a network file system. Materializing the URI to that PDF document fetches the document for viewing by an Analyst.

(U//FOUO) Visualization Integration. The BLACKBOOK2 application is primarily web-based, and currently accommodates interactive visualizations of the data through Java applets.

(U//FOUO) Program Name: Common Ontological Data Environment (CODE)

(U//FOUO) **Sponsoring organization**: Joint Warfare Analysis Center

(U//FOUO) **Performing contractor(s)**: BBN Technologies

(U//FOUO) **Gov't POC Phone Number & E-mail Address**: Gretchen Toliver , (540) 653-3945, gtoliver@jwac.mil

(U//FOUO) **Contractor POC Phone Number & E-mail Address**: John Sumner, (703)284-1232, jsumner@bbn.com

(U//FOUO) **Abstract description**: CODE is a data management architecture based on semantic web technologies that allows analysts to rapidly query and ingest structured data sources, perform deconfliction on their data set, add knowledge and additional information to their data set, and export their reconciled data to Command modeling environments for further analysis and product generation.

(U//FOUO) **Intended users**: JWAC analysts

(U//FOUO) **Catalyst functionality included**:

Entity extraction	
Relationship extraction	
Metadata management	
Semantic entity integration	√
Entity disambiguation	√
Entity knowledge base	√
Visualization	√
Query	√
Knowledge management	√

(U//FOUO) **Sources of input data**: structured data sources to include certain databases, .csv files, and xml files.

(U//FOUO) **Scale of current implementation**: currently scaled to three domain areas of interest.

(U//FOUO) **Status of system**: planning for operational testing.

(U//FOUO) **Where deployed**: JWAC.

(U//FOUO) **COTS/OS/GOTS used**: Based on W3C standards: OWL, SPARQL, SWRL

(U//FOUO) **Size of development effort**:

(U//FOUO) **User experiences**: The capabilities of the architecture have significant potential to shorten data preparation time for modeling and analysis. In some cases, taking the time from months to hours.

(U//FOUO) **Plans for continued development**: Integrate more flexible data importers/exporters, customized deconfliction rule engine, tighter integration with modeling toolsets.

(U//FOUO) Program Name: Future Text Architecture

(U//FOUO) **Sponsoring organization**: Joint Warfare Analysis Center

(U//FOUO) **Gov't POC Phone Number & E-mail Address**: Phil Summerson, (540) 653-6064, psummers@jwac.mil

(U//FOUO) **Abstract description**: The Future Text Architecture is a set of capabilities implemented to facilitate the search and discovery of information in unstructured textual data, and to extract that information using a variety of methods in a structured form. The goal is to help flip the "80-20" (80% of an analyst's time is spent on data prep, 20% on analysis) to "20-80."

(U//FOUO) **Intended users**: JWAC analysts

(U//FOUO) **Catalyst functionality included**:

Entity extraction	√
Relationship extraction	√
Metadata management	√
Semantic entity integration	
Entity disambiguation	
Entity knowledge base	
Visualization	
Query	
Knowledge management	

(U//FOUO) **Sources of input data**: semi- and unstructured textual data.

(U//FOUO) **Scale of current implementation**:

(U//FOUO) **Status of system**: in development.

(U//FOUO) **Where deployed**: JWAC

(U//FOUO) **COTS/OS/GOTS used**: Twister Parallel Data Framework (SMSI), Endeca IAP (Endeca), NetOwl (SRA International).

(U//FOUO) **Size of development effort**: 4 developers: 1 DBA, 2 information management SMEs, 2 analysts (part-time)

(U//FOUO) **User experiences**: Too soon to say, but some prototype feedback with the enhanced search capabilities has been extremely positive.

(U//FOUO) **Plans for continued development**: Working toward IOC in Q4FY08.

(U//FOUO) Program Name: Harmony

(U//FOUO) **Sponsoring organization**: National Ground Intelligence Center

(U//FOUO) **Performing contractors**: CACI, Booz, Allen & Hamilton, Eiden Systems

(U//FOUO) **Gov't POC Phone Number & E-mail Address**: Scott Lawrence, (434) 951-1593; scott.l.lawrence@us.army.mil

(U//FOUO) **Contractor POC Phone Number & E-mail Address**: Roger E. Shropshire, (240) 687-3446; rshropshire@caci.com

(U//FOUO) **Abstract description**: The Harmony Program is the national repository for all media and their related translations in support of GWOT and other national requirements for Document and Media Exploitation (DOMEX) processing and storage. Harmony is responsible for timely dissemination of DOMEX information throughout the IC, DoD, and national law enforcement communities of the United States and key allies.

(U//FOUO) **Intended users**: National Intelligence Community, DoD, National Law Enforcement, Coalition Forces, Tactical Field Commanders.

(U//FOUO) **Catalyst functionality included**:

Entity extraction	√
Relationship extraction	
Metadata management	√
Semantic entity integration	
Entity disambiguation	
Entity knowledge base	
Visualization	
Query	√
Knowledge management	

(U//FOUO) **Sources of input data**: Tatical DOMEX activities worldwide, National Media Exploitation Center, DoD Intelligence Service Production Centers, CIA, DIA, NSA, FBI, DHS.

(U//FOUO) **Scale of current implementation**: Over 1.5 million database records with 50TB of data, documents, media, and translations resident on JWICS, SIPRNET, StoneGhost, and NIPRNET (June '08).

(U//FOUO) **Status of system**: Fully operational on three intelligence networks.

(U//FOUO) **Where deployed**: JWICS, SIPRNET, StoneGhost, and NIPRNET (June '08). Deployable tools deployed throughout the world at all BST's in theater, plus MNFI HQ, Afghanistan, CONUS, and OCONUS sites.

(U//FOUO) **COTS/OS/GOTS used**: COTS: Oracle, SQL Server, Jubliant, Basis Technologies, BBN, Identix, Nexida, Language Weaver, AppTek, and Harmony deployable tools include an automated workflow processor, and a multitude of OCR, Machine Translation, visualization products.

GOTS: Harmony custom software in both the National database and deployable tools

(U//FOUO) **Size of development effort**: Four government and over 100 contractors supporting Harmony daily.

(U//FOUO) **User experiences**: Users are presented with a Google-like user interface for keyword/phrase searches of over 1.5 million database records. They are able to search full-text of English, Arabic, and 50+ other foreign languages. The system allows for a user to save searches and have a "Personal Search Agent" run queries automatically with a daily email report of new/updated records that meet their search criteria. There are over 20,000 multimedia files indexed. The multimedia files are searchable phonetically, textually (Arabic native language text and machine-translated English), facial detection, and key-frames. Advanced search capabilities allow for additional filtering based on key parametric information about each database record. Additional sophisticated search strings can be written to access all metadata elements in the system.

(U//FOUO) **Plans for continued development**: Major areas of planned enhancements surround named entity extraction, additional foreign language search tools, automated metadata creation, support for multibyte foreign language character sets (Unicode), integration with community partners in a SOA environment, expansion to other intelligence networks to support coalition forces

(U//FOUO) **Lessons learned**: Analysts continue to demand DOMEX artifacts to support their GWOT analysis. They require translated data in ever-increasing amounts for link-analysis and terrorist identification. The community does not have sufficient linguists or translators to keep up with the demand for this information. It is imperative that programs such as Harmony leverage existing technological tools, and drive innovated future solutions, that can assist in the triage and categorization of documents and media. The Harmony program is vital to these analytical efforts.

(U//FOUO) Program Name: Information Extraction/Structured Data Analysis (IE/SDA)

(U//FOUO) **Sponsoring organization:** CIA/APPS/Analytic Technology Solutions

(U//FOUO) **Gov't POC Phone Number & E-mail Address:** Gwendolyn G. Graham-Zanin, 703-547-6904, gwendgg@ucia.gov.

(U//FOUO) **Abstract Description:** The Information Extraction/Structured Data Analysis (IE/SDA) project was established to meet enterprise strategic requirements to extract and create structured data from unstructured text.

(U//FOUO) The IE/SDA project has delivered a set of enterprise services that leverage machine-based Natural Language Processing (NLP) processes to identify and extract entities (e.g., people, places, organizations) and relationships from unstructured text. There are two distinct services or systems: (1) bulk processing via a scaleable framework, and (2) on-demand extraction via web services.

(U//FOUO) Both services deliver extraction results in an Agency-approved specification known as the Common Representation Format (CRF). Further transforms can be exacted against the CRF XML standard in order to filter results into formats required by downstream users.

(U//FOUO) Currently, IE/SDA extracts information by request through the on-demand service, as well as runs extraction on a daily basis against documents as they are ingested into the Neptune Data Layer (NDL). These extractions are then made available for use by the enterprise.

(U//FOUO) IE/SDA also works with individual components to meet their specialized business needs for extraction and incorporates the resulting NLP strategies, as needed, into the enterprise extraction processes.

(U//FOUO) **Intended users:** Anyone at CIA in need of extracted information in order to discover and exploit collected intelligence.

(U//FOUO) **Catalyst functionality included**:

Entity extraction	√
Relationship extraction	√
Metadata management	√
Semantic entity integration	√
Entity disambiguation	√
Entity knowledge base	√
Visualization	√
Query	√
Knowledge Management	√

(U//FOUO) **Sources of input data:** Text ingested daily into CIA including message traffic and open source documents. Also text from any source submitted through the on-demand web service.

(U//FOUO) **Scale of current implementation:** Large scale. Currently we have

extracted from over 20 million documents in CIA repositories.

(U//FOUO) **Status of system:** In production since September 2007.

(U//FOUO) **Where deployed:** CIA.

(U//FOUO) **COTS/OS/GOTS used:** SMSi's Twister (scaleable framework); Aerotext, Attensity, ThingFinder, and MetaCarta (extraction); Endeca, Spotfire, In-Spire, Centrifuge, and Palantir (structured data exploitation).

(U//FOUO) **Size of development effort:** approximately 16 FTE.

(U//FOUO) **User experiences:** Positive.

(U//FOUO) **Plans for continued development:** Besides continuing to work with individual components to meet ongoing needs for extraction, we plan to deliver event extraction; document categorization; services for users to create, delete, and enrich extraction results; and services for users to resolve entities across documents.

(U//FOUO) **Lessons learned:**

(U//FOUO) Agile development and 30-day time-blocks are a good thing.

(U//FOUO) Good to work closely with the developers and integrators of analytic tools within the CIA to help users learn how to exploit extracted information.

(U//FOUO) Good to work closely with extraction engine vendors so they know about future capabilities you are looking for.

(U//FOUO) Good to work closely with other extraction groups within the Community (including your own agency) to share knowledge and technologies.

(U//FOUO) Program Name: Intelligence Integration Cell (IIC)

(U//FOUO) **Sponsoring organization**: National Counterterrorism Center (NCTC).

(U//FOUO) **Performing contractor**: The Boeing Team.

(U//FOUO) **Gov't POC Phone Number & E-mail Address**: Vicki J. McBee, vickijm@nctc.gov.

(U//FOUO) **Contractor POC Phone Number & E-mail Address**: Gail Carr, gail.v.carr@boeing.com.

(U//FOUO) **Abstract description**: Analysts in the Information Integration Cell (IIC) bring together data, tools, and methods to perform analysis on information available to the Federal Government that is potentially related to terrorism. Building upon analytic theory and technology using traditional and non-traditional sources of information, data is examined, analyzed, and fused to detect new indications of terrorist activities in a semi-automated process. Insights or leads are provided to the analytical and operational components of the US Counterterrorism (CT) community.

(U//FOUO) **Intended users**: a small cadre of seasoned analyst that support the broader CT mission.

(U//FOUO) **Catalyst functionality included**:

Entity extraction	√
Relationship extraction	√
Metadata management	√
Semantic entity integration	√
Entity disambiguation	√
Entity knowledge base	√
Visualization	√
Query	√

(U//FOUO) **Sources of input data**: databases and document collections from across the IC.

(U//FOUO) **Status of system**: operational.

(U//FOUO) **Where deployed**: NCTC. There is also an unclassified lab located at MITRE for vetting candidate technologies targeted for the IIC.

(U//FOUO) **COTS/OS/GOTS used**: NetOwl, Endeca, Initiate, Oracle, Palantir, Centrifuge, Spotfire, ORA, LLNL's XKE, and more.

(U//FOUO) **Size of development effort**: 30-35 developers, prototypers, support, and embedded technologists, 12 analysts.

(U//FOUO) **Plans for continued development**: continue to bring in new tools and transition them to Railhead when proven.

(U//FOUO) Program Name: KWeb (GeoTASER & Knowledge Miner)

(U//FOUO) **Sponsoring organization:** National Geospatial-Intelligence Agency (NGA).

(U//FOUO) **Performing contractor(s):** Intelligence Data Systems.

(U//FOUO) **Gov't POC Phone Number & E-mail Address:** Gail Naftzger, 703/755-5733, gail.g.naftzger@nga.mil.

(U//FOUO) **Contractor POC Phone Number & E-mail Address:** Brian Meighen, Intelligence Data Systems, 703/755-5617, Brian.E.Meighen.ctr@nga.mil.

(U//FOUO) **Abstract Description:** KWeb is a program that consists primarily of two components: GeoTaser and Knowledge Miner. Kweb is the follow-on effort to the GKB-p program. GKB-p has demonstrated capabilities to support knowledge generation, knowledge management, visualization and information sharing for the NSG. These capabilities are implemented using a Service Oriented Architecture (SOA) approach.

(U//FOUO) Kweb will explore advance technologies and capabilities that will enable the analyst to focus quickly on intelligence issues by having an automated system to retrieve, prepare and present intelligence information at the workstation.

(U//FOUO) The current implementation of GeoTaser operates in the Persistent Surveillance Lab (PSL) in Reston. It is outside of NGAnet, on the NGA portion of JWICS (green space). It operates under a blanket security plan for the PSL, as part of the K-Web effort. The application/services have not been separately accredited outside of the accreditation for the lab to operate as a prototyping facility.

(U//FOUO) The current K-Web ontologies are in RDF, with an expectation of moving to OWL in the future.

(U//FOUO) **Intended users:** IC analysts supporting geospatial-related requirements. Kweb partners are NGA (P, A, GKB, KPE), Mission Partners, and COCOMs.

(U//FOUO) **Catalyst functionality included**:

Entity extraction	√
Relationship extraction	
Metadata management	
Semantic entity integration	
Entity disambiguation	
Entity knowledge base	√
Visualization	√
Query	√
Knowledge management	√

(U//FOUO) **Sources of input data:** Multi-source data types (PDF, text,…). Any document may be processed for extraction of geospatial entities, state-free.

(U//FOUO) **Scale of current implementation:**

(U//FOUO) **Status of system:** Available to the community as prototype.

(U//FOUO) **Where deployed:** Interest has been expressed in the KWeb GeoTASER capability by many agencies within the community including CIA, DIA, DNI, SOCOM, but it is currently not deployed.

(U//FOUO) **COTS/OS/GOTS used:**

(U//FOUO) GeoTASER: InXight, Oracle (10g R2 with spatial and text), GeoNames.

(U//FOUO) Knowledge Miner: InXight, Oracle (plus Thesaurus), TopQuadrant (for ontology development), Saffron.

(U//FOUO) Rules & Alerts: AgentLogic (formerly JRules).

(U//FOUO) Statistical Analysis Visualization**:** SpotFire, Tableau.

(U//FOUO) The application is built to be gazetteer agnostic, plans are to add the USGC GNIS gazetteer for US places, and possibly some specialized gazetteers for countries of interest.

(U//FOUO) **Size of development effort:** 14 people during maximum development.

(U//FOUO) **User experiences:** DIA's Counter Narcotics Division finds the application very useful in support of narco-terrorism requirements.

(U//FOUO) **Plans for continued development:** The application is evolving: the K-Web program is operating under a deadline of 01 October 2008 for transitioning the capability to an operational capability. After that date, its funding for further development ends. They are exploring a number of options:

(U//FOUO) Eventually (between 2009 and 2011), it may be integrated into the Knowledge Management and Mining (KMM), Unstructured Information Management (UIM) portion of the GeoScout Program at NGA. Plans are to turn over the next generation of the code to GeoScout.

(U//FOUO) On a faster track, the Advanced Rapid GEOINT Solutions (ARGS) program in St Louis is looking to field GeoTaser on SIPRNET in the GIAT, for eventual integration into the Gateway.

(U//FOUO) There are no current plans for an unclassified implementation.

(U//FOUO) **Lessons learned:**

(U//FOUO) An early version used MetaCarta. The existing version utilizes the Inxight software development kit and custom code, and utilizes portions of the NGA gazetteer. The entire gazetteer is not utilized for two reasons:

(U//FOUO) Duplicate entries (such as small stream names) that often create false hits.

(U//FOUO) The fact that customizing the Inxight name catalog requires processing long lists in memory, and the entire gazetteer cannot be handled using a single name catalog. (Thus the catalog has been culled, and a distributed name catalog engine that runs in separate Java Virtual Machines on the same server is utilized).

(U//FOUO) A soon-to-be-deployed version will utilize Oracle 10G to manage the gazetteer (bypassing the need for the distributed name catalogs), but making the application run slower (~ 7 seconds per document, as opposed to 1-2 seconds per document with the other approach). Developers are still tweaking the performance of this implementation.

(U//FOUO) Regarding functionality: Some limited testing suggests that recall (does the tool find all place names?) may be between 70-85% of that of MetaCarta, probably explained by the fact that right now it draws primarily from GeoNames (and a culled down version of it) as opposed to MetaCarta's larger gazetteer. GeoTaser claims that precision (is it really a place, as opposed to a person name, and did I disambiguate the place correctly?) is actually better than MetaCarta, due to the use of Inxight's NLP capabilities to eliminate proper nouns that could be places or other things (e.g. people, organizations), and due to algorithms added for disambiguating given the context of the place name in a sentence. The tests on this were very informal, over a short period of time.

(U//FOUO) Program Name: LSIE = Large Scale Internet Exploitation Project

(U//FOUO) **Sponsoring organization**: DNI Open Source Center

(U//FOUO) **Performing contractor(s)**: L3 Communications

(U//FOUO) **Gov't POC Phone Number & E-mail Address**: Laura Knudsen, 703-613-5917, laurak@rccb.osis.gov

(U//FOUO) **Abstract description**: LSIE is developed as a service-oriented architecture for discovery, ingestion, storage, and processing of open source data. It is largely a COTS/GOTS integration effort, based on open APIs and standards. It is the Open Source Center's new capability to exploit massive amounts of data on the Internet in support of DNI's missions.

(U//FOUO) LSIE takes as input Internet documents and web pages found by either crawling or targeted querying. It also ingests OSC products and can ingest any unclassified data. The data goes through an ingest process that includes language identification and entity extraction, as well as indexing. The resultant data is marked up in XML and stored in a specialized database optimized for storage of XML documents. Portals, query tools, and APIs then access the database. A "knowledge layer" and machine translation are also included. In the future, analytical tools and alerting/profiling against the data will be supported. Access is via thin client (web browser). Service management and security capabilities are built into the infrastructure on which LSIE is developed.

(U//FOUO) LSIE Description:

- Massive volume of unstructured, multilingual, multimedia open source data
- Management via taxonomies
- Open APIs into repository and evolving set of analytic tools allows mining of diverse data pool
- Knowledge sphere: human-machine interaction builds new corpus of intelligence data available to all

(U//FOUO) **Intended users**: Entire IC (not just OSC).

(U//FOUO) **Catalyst functionality included**:

Entity extraction	√
Relationship extraction	
Metadata management	√
Semantic entity integration	√
Entity disambiguation	√
Entity knowledge base	√
Visualization	√
Query	√
Knowledge management	√

(U//FOUO) **Sources of input data**: Crawling and searching the Internet.

(U//FOUO) **Scale of current implementation**: Over 500M resources (multilingual documents) will be in LSIE by August 2008.

(U//FOUO) **Status of system**: Functional prototype. IOC planned for September 2008.

(U//FOUO) **Where deployed**: OSC.

(U//FOUO) **COTS/OS/GOTS used**: BrightPlanet for crawling and deep content identification; Basis for language identification, CyberTrans and Language Weaver for machine language translation, Oracle for targeting database, Stellant and InXight for information extraction (InXight on 9 languages), MarkLogic for storage and knowledge layer (includes knowledgebases created using OSC SpyGLAS efforts). Prototyping is being done using Metacarta, Prefuse, Tibco Spotfire and FMS Sentinel for visualization.

(U//FOUO) **Size of development effort**: ~60 FTEs.

(U//FOUO) **User experiences**: An early version (Sept. 07) showed high value but usability issues. Participants in alpha testing were from across the IC. These issues have been addressed in the current version.

(U//FOUO) **Plans for continued development**: Get to IOC, support operational use, work with community members who want to use the system and/or APIs, continue enhancing.

(U//FOUO) **Lessons learned**: 1) Analyst involvement is critical for success.

2) There must be a balance between strategic and tactical goals.

3) Multiple niche skills are necessary for implementation.

4) Requirements management for a COTS/GOTS integration effort is different from that for a custom development effort.

(U//FOUO) Program Name: Metadata Extraction and Tagging Service (METS)

(U//FOUO) **Sponsoring organization:** Defense Intelligence Agency/Enterprise Services (DIA/ES)

(U//FOUO) **Performing contractor(s):** BAE (prime) with subs Booz Allen Hamilton, InXight, others

(U//FOUO) **Gov't POC Phone Number & E-mail Address:** Tim Giles, 202/231-3814, Timothy.Giles@dia.mil

(U//FOUO) **Contractor POC Phone Number & E-mail Address:** Mel Laney/BAH, 703-981-7720, laney_melvin@bah.com.

(U//FOUO) **Abstract Description:** METS is a DIA core service and data infrastructure component that automatically extracts entities and the semantic relationships among them from unstructured documents through sophisticated dissection and knowledge-engineered automated procedures. METS was developed in support of the Defense Intelligence Agency Strategic Plan (Fiscal Years 2004–2009).

(U//FOUO) METS provides a central metadata tagging and entity extraction "factory" for use by IC applications and portals. The resulting RDF triples (in Web Ontology Language (OWL) and XML) provide support to virtually any application or user interface required. METS packages several COTS tools along with the necessary knowledge engineering and interfaces to provide a centralized IC data engine, thereby alleviating high costs for individual organizations to create a like capability. METS includes the ability to: extract persons, organizations, locations and other entities, and events, from collection sources, finished intelligence, and specific open source materials; normalize the documents into a standard format; tag entities using XML; extract semantic information such as properties and relationships; integrate views of tagged data; and create and deliver appropriate information to end-users through a variety of applications, portals, and knowledge bases. This functionality significantly enhances the ability of analysts to quickly search and merge data from databases and data sources throughout the community.

(U//FOUO) The principal forms of output from METS include XML "dialects"; i.e., XML and RDF/OWL. Tags used in the outputs are based on a highly generalized ontology developed to support analysts and augmented with intelligence domain specific sub-classes, properties, and tags. In addition, technical ontologies are used to augment the general ontology where appropriate, e.g., IC metadata standard for publication (ICMSP) metadata elements.

(U//FOUO) Earlier versions of METS included a persistent knowledge base of extracted entities, but in the next version (3.0) METS will be only a stateless, on-demand web service. Persistent storage is done in other parts of the DoDIIS architecture.

(U//FOUO) **Intended users:** Intelligence Community (IC) agencies and members.

(U//FOUO) **Catalyst functionality included:**

Entity extraction √

Relationship extraction	√
Metadata management	√
Semantic entity integration	
Entity disambiguation	√
Entity knowledge base	
Visualization	√
Query	√
Knowledge management	

(U//FOUO) **Sources of input data:** Collection sources, finished intelligence, and specific open source materials. (U//FOUO) **Scale of current implementation:** METS currently can process 50-60K documents per day. This has been determined to be insufficient, hence the move to multithreaded implementation. The goal is 250K documents per day.

(U//FOUO) S**tatus of system:** Currently in Version 2.5, testing for operational use. For 3.0, implementation is on 8 CPU machine for multithreaded application (single thread was too slow). There are two systems, one for legacy data and one for real time needs. After the legacy data is processed, that system will move to SIPRNet. Currently accredited to PL3.

(U//FOUO) **Where deployed:** DIA on JWICS. It is available to any organization within the SYSNET2 SLA governance structure.

(U//FOUO) **COTS/OS/GOTS used:** Oracle 10g with Spatial Extensions (moving to other project in 3.0), InXight Suite (particularly ThingFinder), AeroText, and Attensity for extraction (the latter two will be dropped in 3.0).

(U//FOUO) **Size of development effort:** Approx. 4 FTEs.

(U//FOUO) **User experiences:** Performance achieved: ~80-90% for entity extraction, ~60% for relationships.

(U//FOUO) **Plans for continued development:** Enhanced web services architecture. Accreditation to PL4 planned.

(U//FOUO) **Lessons learned:** Earlier version had Tucana for storage, but it had reliability issues, so Oracle replaced Tucana. Also, DIA has an enterprise license for Oracle, and accreditation is easier with Oracle.

(U//FOUO) Program Name: Pathfinder

(U//FOUO) **Sponsoring organization**: NGIC/ES.

(U//FOUO) **Gov't POC Phone Number & E-mail Address**: Dave Patterson, 434-951-1803, david.k.patterson1@usarmy.mil.

(U//FOUO) **Abstract description**: Pathfinder is a web enabled capability that provides analysts with search and discovery tools that allow them to perform the following analytic functions:

1) Data harvesting tools to collect, normalize, and extract and tag entity and geo-references; users have the ability to create and apply lists of entities for custom extraction.
2) Use Boolean logic queries to return high-precision results from a large collection of intelligence reporting (archive back to 1988).
3) Query building tools such as sounds-like, wildcard matching, entity alias lists, query by example, and others.
4) Apply a range of tools on search results to perform trend and pattern analysis on large sets of reporting.
5) Automatically establish link-diagrams and geo-plot overlays from specific search results with click-able links back to the originating intelligence report.
6) Perform geographic based searches on a map area bounded by a box, polygon or line/route on all collected intelligence reporting (including message traffic and tactical reporting collections).
7) Collaborate with other users to share queries/search models and vetted data collections.

(U//FOUO) **Intended users**: Intelligence analysts from tactical to strategic. Current implementations are well suited for All-Source/GMI, HUMINT, and S&T analysts.

(U//FOUO) **Catalyst functionality included**:

Entity extraction	√
Relationship extraction	√
Metadata management	√
Semantic entity integration	√ (non-automated analyst-driven tools)
Entity disambiguation	√
Entity knowledge base	√ (data catalog available on a portal)
Visualization	√
Query	√
Knowledge management	

(U//FOUO) **Sources of input data**: Well formatted text (M3, WISE), structured databases, unstructured data (HTML/Word/PowerPoint) from a wide range of sources. Complete list is classified.

(U//FOUO) **Scale of current implementation**: ~30 implementations currently in place throughout the world, ranging in user-base of 10-1000 users at each site. Fielded on five

US / coalition networks, and one foreign network (UK). Established presence at DoD locations and in an enterprise implementation available to all SIPRNet and JWICS users.

(U//FOUO) **Status of system**: Operationally fielded and maintained at sites either with local administrative staff or remotely from a central location.

(U//FOUO) The Pathfinder project office at NGIC concluded development efforts in Aug 2007 at the close of a contract. Further software development and maintenance are being managed by INSCOM Futures. Current development efforts are focusing on integration with the Army's Distributed Common Ground System (DCGS-A). An INSCOM/NGIC Task Force is in place to manage the transition to the combined system, and tie into other initiatives lead by INSCOM.

(U//FOUO) **Where deployed**: Multiple locations on multiple networks.

(U//FOUO) **COTS/OS/GOTS used**:

COTS: Memex (search engine), custom Lucene syntax translator (SAIC)

OS: Lucene (search engine)

GOTS: Pathfinder analytic and datamining tools

(U//FOUO) **Size of development effort**: N/A

(U//FOUO) **User experiences**: Generally favorable, but some don't prefer the capability as it stands in a web browser. The DCGS-A Multi-Function Workstation (MFWS) is a newly developed windows look-and-feel interface to the Pathfinder search tools.

(U//FOUO) Some users have also expressed difficulties with the query syntax. An added "fielded" search functionality is now available to accommodate users who don't need the precision of Boolean logic query syntax.

(U//FOUO) **Details**: At data-load time, data mapping, entity extraction and geo-location extraction are performed. Data mapping is a process that processes well formatted data sources and normalizes the data types into a consolidated tagging methodology.

(U//FOUO) Entity and geo-location extraction finds and tags entities by either matching with a list of entities or through regular expressions. The entities currently tagged in Pathfinder data sources are:

- BE Number
- Date
- Relative date
- Person (Western and Arabic)
- Organization
- Country
- Equipment/Weapons
- Facility
- Military Unit
- Telephone number
- IP address
- Email address
- URL

(U//FOUO) Geocoordinate data currently tagged and normalized to MGRS are:

- UTM/MGRS
- LAT/LON (degree minutes seconds)
- Decimal Degrees

(U//FOUO) Program Name: Quantum Leap

(U//FOUO) **Sponsoring organization**: CIA.

(U//FOUO) **Performing contractor(s)**: White Oak Technologies, Oracle, L3 Communications, others.

(U//FOUO) **Gov't POC Phone Number & E-mail Address**: William Haynes, 703/547-0566, williph0@ucia.gov.

(U//FOUO) **Abstract description**: Quantum Leap (QL) discovers knowledge in massive, disparate intelligence data to address national intelligence priorities with the CIA through excellence and innovation in data aggregation, tools, analytic methods, and dissemination.

(U//FOUO) QL combines intelligence data, proven commercial technologies and analytic expertise to:

1. find non-obvious linkages, new connections, and new information from within the data, and
2. acquire and integrate additional intelligence data that improves the value of the current integration.

(U//FOUO) QL continually seeks out, proves, and applies new data technologies to intelligence issues.

(U//FOUO) QL is poised to impart its data discovery technology on the CIA Enterprise Data Layer.

(U//FOUO) **Intended users**: Supporting 20 CIA organizations (146 branches), primarily NCS and CTC.

(U//FOUO) **Catalyst functionality included**:

Entity extraction	√
Relationship extraction	√
Metadata management	√
Semantic entity integration	
Entity disambiguation	√
Entity knowledge base	
Visualization	√
Query	√
Knowledge management	

(U//FOUO) **Sources of input data**: Classified.

(U//FOUO) **Scale of current implementation**: 10 terabytes of raw data, many more processed, 1000s of CPUs working problem space in parallel.

(U//FOUO) **Status of system**: Operational since 2003.

(U//FOUO) **Where deployed**: CIA.

(U//FOUO) **COTS/OS/GOTS used**: White Oak Technologies, Inc. (WOTI) Wareman for entity disambiguation, Lexis Nexus Data Supercomputer for high speed query support, Netezza for high speed query support, Netowl, and Serotexct for entity extraction, Centrifuge for visualiztion, ESRI suite for visualization of GIS data, QLIX (GOTS) for display of disambiguation results, Plasma (GOTS) for pedigree and lineage metadata tracking. Formerly used NORA (too slow to load on hardware, doesn't work well on sparse data), Initiate (didn't have the staff to experiment with), and Attensity (the early release was too intensive to train).

(U//FOUO) Size of development effort: ~30 FTEs.

(U//FOUO) User experiences: QL has proven valuable in producing a variety of products resulting from simple searches or a far more complex analysis.

(U//FOUO) Plans for continued development: QL continues its interest in entity resolution. QL is attempting to develop algorithms and/or indices of data to support broader operations. For the first time, QL is shifting focus to the presentation layer and attempts to provide analysts with better ways to organize their information.

(U//FOUO) Lessons learned: (1) Be prepared to do everything over multiple times, until it is "correct" (and "correct" will change over time, so you need to always be prepared to reprocess all data). This principle affects every aspect of the program, from the hardware to the sizing to the labor effort to the scheduling to … (2) Different analysts want different analytical processing of the data. Getting agreement on even the simplest of forms is difficult. For example, the QL date format is standardized (YYYYMMDD), with a set of rules for what to do when the data is invalid in some way (like 31 Feb), but certain analysts wanted entity resolution done on the "raw" format of the date, to take advantage of patterns of similar errors.

(U//FOUO) Program Name: SAVANT (Systematic Architecture for Virtual Analytic Net-Centric Threat Information)

(U//FOUO) **Sponsoring organization:** National Air and Space Intelligence Center (NASIC)/Advanced Programs Directorate.

(U//FOUO) **Gov't POC Phone Number & E-mail Address:** Dan Geragosian, Program Manager, 937-257-5100, Daniel.Gerogosian@WPAFB.AF.Mil.

(U//FOUO) **Abstract Description:** One of NASIC's major accomplishments in Information Sharing is the Systematic Architecture for Virtual Analytic Net-Centric Threat Information (SAVANT) – a Service-Oriented corporate architecture that enables documenting, storage, and presentation of corporate knowledge in a standard manner. SAVANT allows analysts to define what data they want to store, how the want to share it, and how to present a product which can be shareable to the community.

(U//FOUO) **Intended users:** To be installed: AFIWC, ONI, NGIC, 53TW, China Lake, Pt Mugu TC. Evaluating for use: AFSPC, CDP, JCS J2J, DIA-DI JWS, 480th IW.

(U//FOUO) **Catalyst functionality included**:

Entity extraction	√
Relationship extraction	√
Metadata management	√
Semantic entity integration	√
Entity disambiguation	√
Entity knowledge base	√
Visualization	√
Query	√
Knowledge management	√

(U//FOUO) **Sources of input data:** Multi-Source INTs.

(U//FOUO) **Status of system:** Operational. Domain specific deployments in works.

(U//FOUO) **Where deployed:** NASIC and MSIC.

(U//FOUO) Program Name: VICTORE (Vocabularies for the IC to Organize and Retrieve Everything)

(U//FOUO) **Sponsoring organization**: CIA/CIO/APPS/DAS.

(U//FOUO) **Performing contractor(s)**: various under I2S.

(U//FOUO) **Gov't POC Phone Number & E-mail Address**: Kevin Lynch, 703-613-8815, kevinsl@ucia.gov.

(U//FOUO) **Contractor POC Phone Number & E-mail Address**: Michael Hudson, 703-613-8837, Mike.Hudson@ngc.com.

(U//FOUO) **Abstract description**: VICTORE (Vocabulary for IC to Organize and Retrieve Everything) is to develop an "Intelligence Topics Controlled Vocabulary" (ITCV) to describe the terms used for subject matter and other metadata associated with intelligence documents. It will provide controlled vocabulary based on a logical data model to assure integrity of the relationships among terms. The result will be a neutral formal vocabulary, taxonomy and set of master data instances that covers the same semantic area as current conventions, and will be mapped to current conventions, but is not tied to any one convention. It will isolate systems and users from changes in tagging and markup conventions. It will also form a firm foundation for query enrichment and taxonomy mapping.

(U//FOUO) **Catalyst functionality included**:

Entity extraction	
Relationship extraction	
Metadata management	√
Semantic entity integration	
Entity disambiguation	
Entity knowledge base	
Visualization	
Query	
Knowledge management	√

(U//FOUO) **Sources of input data**: Existing conventions such as NIPF, IFC, target and other encoding standards, subject matter experts and other reference material.

(U//FOUO) **Scale of current implementation**: Small (pilot) user population.

(U//FOUO) **Status of system**: Pilot

(U//FOUO) **Where deployed**: JWICS

(U//FOUO) **COTS/OS/GOTS used**: Knoodl wiki tool, built by Revelytix supplemented with tools and databases.

(U//FOUO) **Size of development effort**: One developer, with advisory committee.

(U//FOUO) **User experiences**: not a significant population of users to date.

(U//FOUO) **Plans for continued development**: Have developed a process for analysis of existing topic-oriented labeling schemes and synthesis of formalized concepts as the

foundation for the ITCV. The ICTV concepts would then be mapped to Terms used in other conventions and to Master Entities in a database. Mappings would be exposed, assessed and improved by SMEs and IMOs. Since the participation and cooperation of these people is vital, a solid foundation must be in place before approaching them to avoid the perception of wasting their time.

(U//FOUO) **Lessons learned**: Useful controlled vocabulary and mappings is a complex area that requires careful consideration and integrity of constructs, and must be based on a solid logical model with sufficient rigor and integrity to support enterprise services for management and dissemination. This is a new area into which a lot of effort has already been poured, some of it shortsighted and unproductive. Any significant progress must be based on collaborative effort, continuity and trust.