IMT: A Mixed-Initiative Data Mapping and Search Toolkit

Michael Zang[1], Adam Gray[2], Joe Kriege[1], Kalyan Moy Gupta[3], David W. Aha[4]

[1]CDM Technologies, Inc.; San Luis Obispo, CA 93401
[2]Collaborative Agent Design Research Center (CADRC);
California Poly Technical State University; San Luis Obispo, CA 93405
[3]Knexus Research Corp.; Springfield, VA 22153
[4]Navy Center for Applied Research in Artificial Intelligence;
Naval Research Laboratory (Code 5514); Washington, DC 20375
{mzang,adgray,jkriege}@cdmtech.com kalyan.gupta@knexusresearch.com david.aha@aic.nrl.navy.mil

Abstract

Interoperability requires the resolution of syntactic and semantic variations among system data models. To address this problem, we have developed the Intelligent Mapping Toolkit (IMT), which employs a distributed multi-agent architecture to enable mixed-initiative mapping of metadata and instances. This architecture includes a novel federation of service-encapsulated matching agents that leverage case-based reasoning methods. We have recently used the IMT matching service to develop several domain-specific search applications in addition to the IMT mapping application.

The Motivation for Developing IMT

Interoperability among information systems is a primary concern in integrating processes both within and across organizations. As the distribution process owner (DPO) for the U.S. Military, this is particularly true for the United States Transportation Command (USTRANSCOM), which integrates distribution processes (e.g., supply requisition, inventory management, and transportation) across the individual military services, suppliers, shippers, and host nation support systems. To facilitate the requisite levels of interoperability among system-specific information models, USTRANSCOM has developed the Distribution Process Information Exchange Data Model (DPIEDM) and initiated an effort to map existing system-to-system interfaces to this logical data model. DPIEDM's goal is to provide a much improved semantic and contextual specification to information exchanges, thus improving current and future process integration across the extended enterprise.

The essential operation in data mapping is *Match*, which takes two schemas (or table extensions) as input and produces a mapping between elements of them that correspond semantically (Rahm and Bernstein 2001). For two schemas with n and m elements respectively, the number of possible matches is $n*m$, implying a manually prohibitive effort when mapping to schemas containing thousands of elements, such as the DPIEDM. This implication prompted USTRANSCOM to automate aspects of their mapping task to significantly decrease the requisite level of effort (i.e., time and expertise) while reducing errors. No usefully-applicable, commercial products for semantic mapping automation exist. Thus, USTRANSCOM sponsored the development of the IMT operational prototype, which applies Artificial Intelligence (AI) techniques to this compelling problem. The IMT project was a collaborative endeavor involving CDM, CADRC, Knexus, and NRL, and USTRANSCOM's semantic mapping community.

The IMT Prototype Description

We introduced IMT in our IAAI-08 paper *Enabling the Interoperability of Large-Scale Legacy Systems* (Gupta, et. al. 2008). Here we summarize it only briefly. IMT proves novel in several ways. It maps large-scale schema (i.e., metadata) and instance data. It employs a distributed multi-agent architecture that includes a federation of matching agents for case-based similarity assessment and learning. IMT semi-automatically acquires domain-specific lexicons and thesauri to improve its mapping performance. It also provides an explanation capability for mixed-initiative mapping. ***IMT's primary goal is to suggest mappings to users for final verification and acceptance***. Its architecture includes the three layers of components shown in Figure 1 and described below.

The ***GUI Layer*** comprises a graphical user interface that allows users to perform actions such as importing, selecting, and visualizing problem elements; acquiring auxiliary resources; invoking matching agents; consulting the agent explanation facility; and exporting mapping solutions for use in other applications.

The ***Agent Layer*** provides *Matching* agents that compute the similarity between problem elements (i.e., tables and fields) by employing similarity assessment procedures typically used in case-based reasoning (CBR). Each agent uses a different feature representation to address a variety of syntactic and semantic variations. For example, the N-gram Matcher converts element names and descriptions into n-grams, each of which becomes a feature, to address the morphological variations in the text pertaining to verbs and nouns (e.g., description vs. describe). Likewise, the Word Matcher tokenizes multi-word descriptions into words that will be used as features

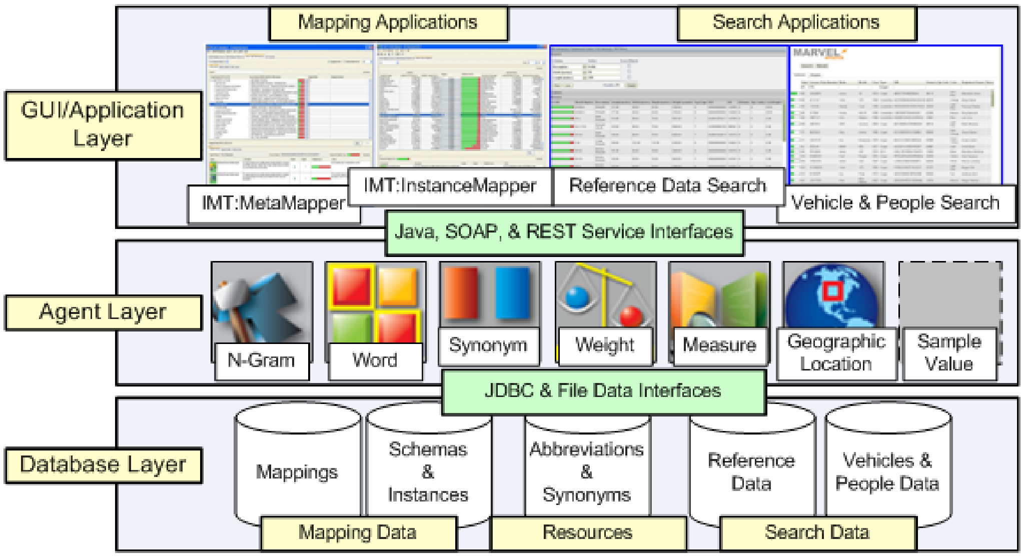


Figure 1: The IMT Architecture

for linguistic matching. Unlike the N-gram Matcher, the Word Matcher users inputs from the Synonym Matcher to process semantic variations. The Synonym Matcher computes the similarity of two features by using the Abbreviations and Synonyms Libraries. The Word Matcher then incorporates these results into the overall similarity assessment.

The ***Database Layer*** includes JDBC-compliant repositories for persisting the mapping problem and solution representation—supporting mapping among schemas, tables, fields, and instances—and the resources for storing the abbreviations and synonyms—supporting the strength of association among synonyms for use by matching. Additionally, schema and instance data may be imported directly from mapping problem sources.

The New Capabilities and Applications

Since completion of the initial IMT prototype for USTRANSCOM, the underlying similarity assessment framework and agents have been re-factored and cleanly partitioned into a Similarity Assessment Service supporting a number of interfaces (e.g., Java, SOAP, and REST), and a new IMT semantic data mapping toolkit revision. This approach has generalized the original GUI Layer into an Application Layer supporting other problem domains.

In addition to the IMT application, the Similarity Assessment Service now supports domain-specific search tools including: (1) an application to identify desired records in military reference data, and (2) an application to identify vehicles or people of interest. Capabilities currently under development for the IMT mapping application include an agent providing schema match scores from corresponding sample data values, and the generation of data transformation code from the semantic mappings produced by IMT.

The Demonstration

This demonstration employs a combination of display posters, self-running slide-shows, hands-on software interaction by attendees, and narrated software presentations to show the ability of IMT to specify, import, and refine a metadata or instance data mapping problem. The demonstration further illustrates the practical decision support assistance provided by the IMT towards resolving these problems. Additionally, our demonstration will incorporate one or more intuitive IMT-technology-derived search applications developed as Cal Poly student senior projects under the auspices of the CADRC[2]. These applications will show the weighted combination of multiple *Match* methods—including N-Gram, Word with Synonym replacement, Measured Quantity, Geographic Location, and Sample Value comparison—to assess similarity between distinct data elements such as the schemas, tables, and fields of two databases to be mapped.

References

Gupta K.M., Aha D.W., & Moore P.G. (2006). Rough set feature selection algorithms for textual case-based classification. *Proceedings of the Eighth European Conference on Case-Based Reasoning* (pp. 166-181). Ölündeniz, Turkey: Springer.

Gupta, K.M., Zang, M.A., Gray, A., Aha D.W., Kriege J., (2008). Enabling the Interoperability of Large-Scale Legacy Systems. *IAAI-08 accepted paper, Submission Track: Emerging Application or Methodologies Papers*.

Rahm, E., & Bernstein, P.A. (2001). A survey of approaches to automatic schema matching, *The VLDB Journal*, 10, 334-350.

Demo Summary

This demonstration employs a combination of display posters, self-running slide-shows, hands-on software interaction by attendees, and narrated software presentations to show the ability of Intelligent Mapping Tool (IMT) to specify, import, and refine a metadata (i.e., schema) or instance data (i.e., record) mapping problem. The demonstration further illustrates the practical decision-support assistance IMT provides towards resolving these problems. Additionally, our demonstration will incorporate intuitive IMT-technology-derived search applications developed as student senior projects under the auspices of the California Polytechnic State University Collaborative Agent Design and Research Center (CADRC). These applications will demonstrate the combination of multiple data matching methods to assess similarity between distinct data elements. Semantic methods include the statistically-weighted comparison of n-grams and words—with synonym replacement. Numeric methods include comparison of measured quantities and geographic locations. Data elements may correspond to schemas, tables, fields and records in the *Mapping* problem, or a query and database for *Search*.

Demo Storyboard

This document describes the representational elements of two distinct software demonstrations, the IMT Mapping Tools and the IMT Search Tools, proposed for demonstration at IAAI-08.

IMT Mapping Tools

Overview

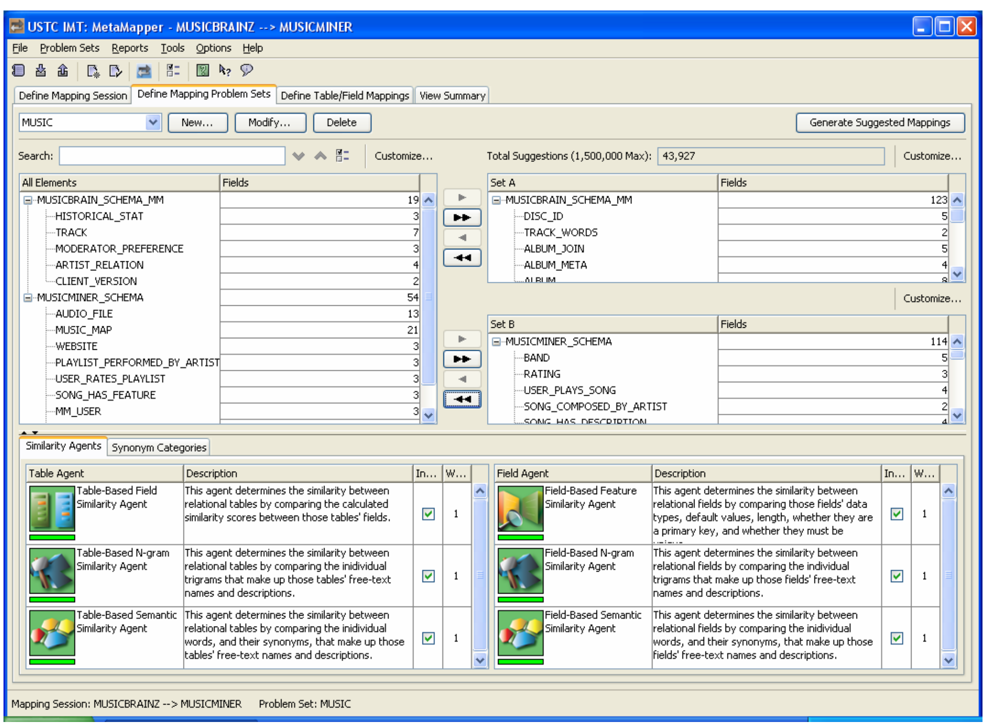
Analysts for the MusicMiner software application desire interoperability with the MusicBrain system, a popular application with an overlapping domain. To support the creation of semantic mappings between the two systems' underlying schemas the analysts have turned to the Intelligent Mapping Toolkit (IMT).

Step 1

The analyst imports xml-based schemas for both MusicMiner and MusicBrain into IMT and selects the tables of interest are from each. He also selects the similarity agents to use for mapping suggestion generation providing the following confidence factors for each:

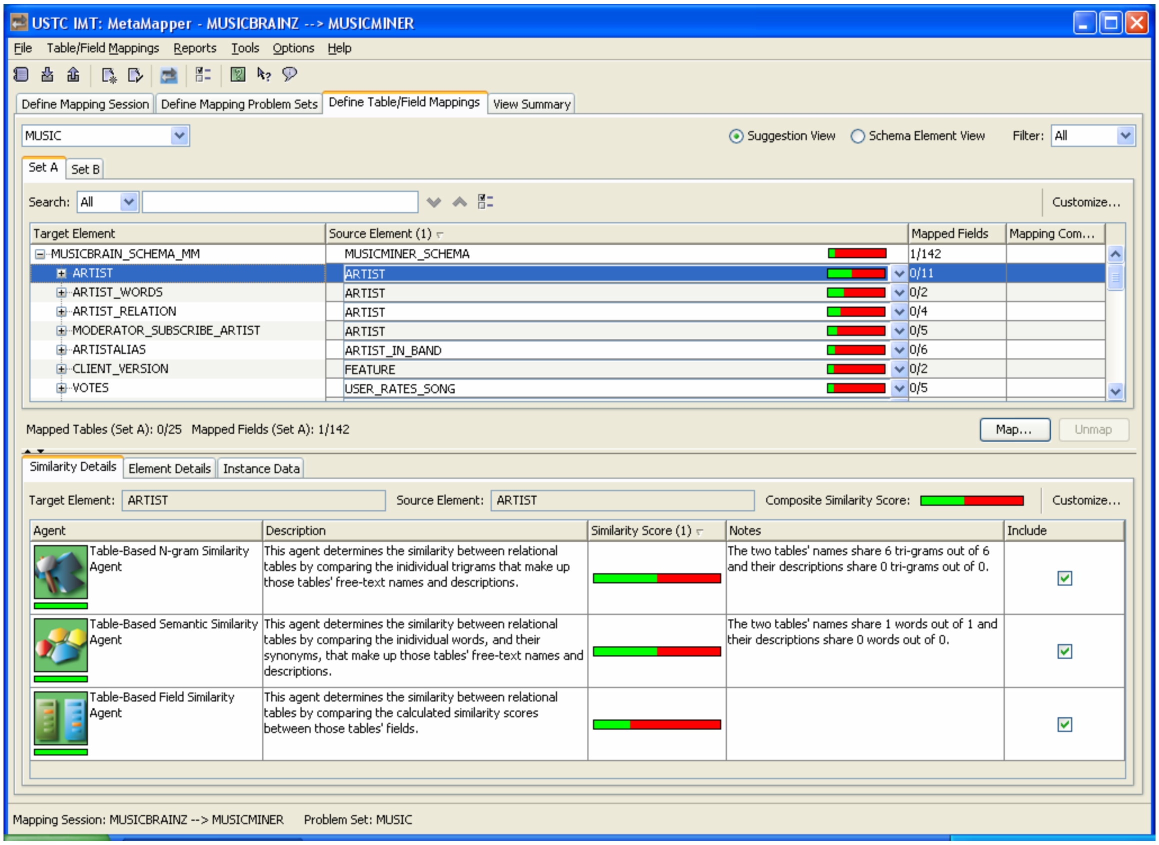
- Feature Similarity Agent (1.0)
- N-gram Similarity Agent (0.5)
- Semantic Similarity Agent (0.5).

The analyst then clicks the "Generate Suggested Mappings" button.



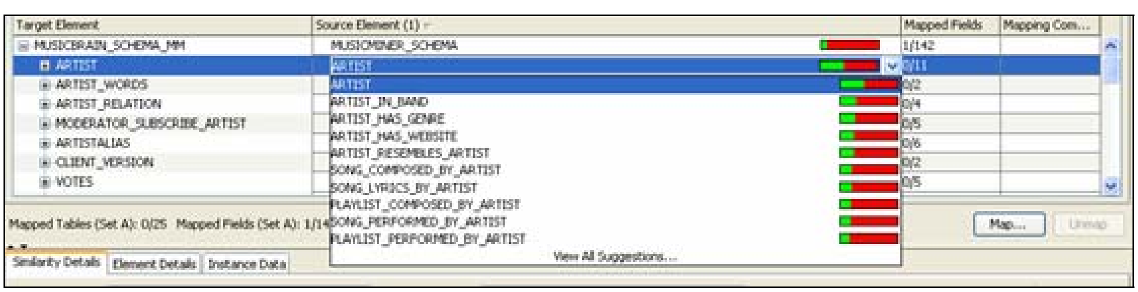
Step 2

The analyst selects the "Define Table/Field Mappings" tab. Suggested mappings for the elements of the MusicMiner and MusicBrain schemas are displayed in the main pane and ordered by similarity score.



Step 3

The analyst expands the combo box located by the "ARTIST" field in the MusicBrain schema. The top 10 most similar fields to "ARTIST" from the MusicMiner schema are displayed. The analyst selects the "ARTIST_IN_BAND" element from the combo box and clicks the "Map…" button.

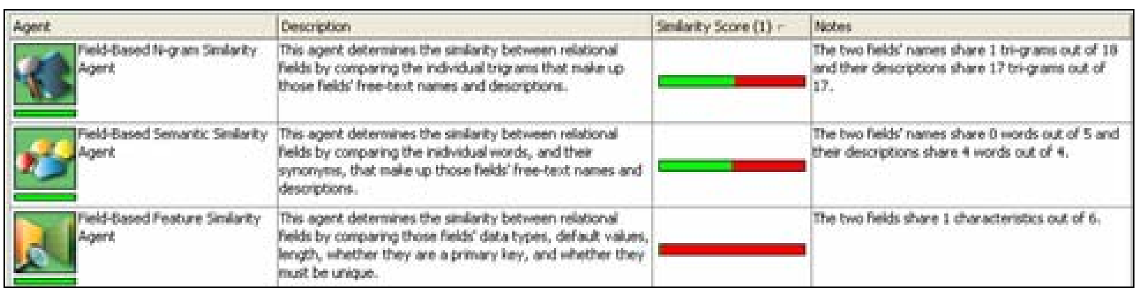


Step 4

The analyst expands the “TRACK” table under the MusicMiner schema to display its associated fields. The analyst notices that the “TRACK_GID” field’s highest ranked suggested mapping is to the “SONG_GROUP_ID” field from the MusicBrain schema.

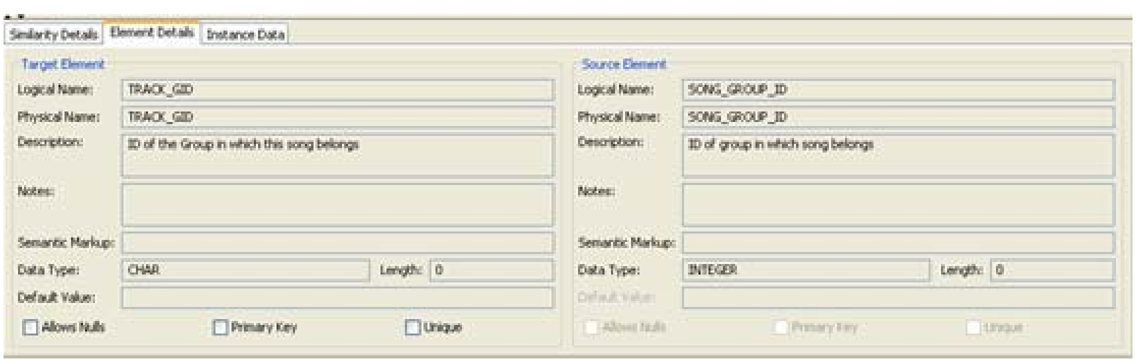
Step 5

The analyst selects the “TRACK_GID” and “SONG_GROUP_ID” row in the schema table. The individual similarity agent results are displayed in the pane below and the analyst notices that both the n-gram and semantic similarity agent results are near 50% despite a strong disparity in names. Upon further inspection, he sees that while the two elements names are entirely dissimilar, their descriptions share 4 words out of 4.



Step 6

The analyst selects the “Element Details” tab and notices that the elements share near identical descriptions.



Step 7

The analyst closes the IMT application.

IMT Search Tools

Overview

A police officer arrives at the scene of a hit and run and interviews an eyewitness who tells him the following information about the suspect's vehicle:

- it was a green SUV,
- it was relatively new, and
- it had a license plate number ending in '4A'.

Step 1

The next day at the police station, the officer attempts to compile a list of potential suspects. He starts by opening the Marvel Search web application.

Step 2

The officer types 'SUV' into the *Type* field and '4A' into the *License Plate Number* field, then clicks the *Search* button.

State	License Plate Number	Make	Model	Year	Type	VIN	Owner's Zip Code	Color	Registered Owner	Notes
	4A				SUV					
NV	MMAD4A	Chevrolet	Tracker	2003	SUV	X9MFGOND1LFSPZ5JO	89502	Dark Brown	Aksel Brazil	
IL	0AE6P4A	Geo	Tracker	1996	SUV	P4R5HOV2WYPPEZXS4	62376	Olive	Mercade Bodo	
MI	R4E864A	Jeep	Grand Cherokee	1994	SUV	674ATIGN6CZU9HH4M	49093	Khaki	Laina Tamir	
NJ	4ACB8BI	Chevrolet	Tracker	1998	SUV	B0H0KV91093CODO0R	8872	Magenta	Adony Amadi	
NC	P8R2NZJ	Buick	Rainier	2007	SUV	GC5GASQVSKA5SRFZX	28644	Light Brown	Curt Malachi	
IN	0H391W	Acura	MDX	2004	SUV	TN0S5E3RO5ZJRVATX	46348	Dark Orange	Orrin Cailean	
MI	O3H2PG	Honda	CR-V	2005	SUV	5WPXMIR0H3WJ0GHGH	49874	Dark Orange	Bayley Amos	
FL	2IZ0JX	GMC	Yukon	1995	SUV	PPN315FTV2JZZP0IS	33830	Violet	Cutter Rurik	
DE	W82PKZX	Ford	Explorer Sport	2001	SUV	T0QKAL4IH7BTKVQYB	19713	Blue	Timeus Roland	
AR	PH0TRT	HUMMER	H1	2006	SUV	3KV423IFTTWMO0PHD	72339	Green Yellow	Mandek Beval	
FL	9WEZA6	Nissan	Xterra	2005	SUV	KI0AF0JG5Y63OKYQE	32834	Blue	Shulamith Orsen	
NE	6SHW4R	Ford	Excursion	2004	SUV	AWQRCD0V2NRN9S1Q4	69356	Silver	Darcy Linh	
TX	I2FUAO	Lexus	GX 470	2003	SUV	44P65DA4H4AJTXR2Y	76452	Pink	Dunbar Axton	
MN	VYT8Q7	Honda	Passport	2000	SUV	BL2MU05RCLQXL02YI	56160	Hot Pink	Lukina Dionysus	
IL	GNPGYC	GMC	S15 Jimmy	1991	SUV	R38TE34V8MQYYHQVT	61644	Violet	Merlin Alwyn	

Step 3

The officer looks at the results returned from his search and sees that there are several SUVs with a license plate ending in '4A'. He clicks on the green score bar for the first result and sees that its combined score is 60.1%, resulting from a 100% match on *Type* and a 20.3% match on *License Plate Number*.

	State	License Plate Number	Make	Model	Year	Type	VIN
		4A				SUV	
60.1%	NV	MMAD4A	Chevrolet	Tracker	2003	SUV	X9MFGOND1LF
	[illegible]	[illegible]	Geo	Tracker	1996	SUV	P4R5HOV2WYI
	[illegible]	[illegible]	Jeep	Grand Cherokee	1994	SUV	674ATIGN6CZU
	NJ	4ACB8BI	Chevrolet	Tracker	1998	SUV	B0H0KV91093C
	NC	P8R2NZJ	Buick	Rainier	2007	SUV	GC5GASQVSK
	IN	0H391W	Acura	MDX	2004	SUV	TN0S5E3RO5Z

Type	100.0%
License Plate Number	20.3%

Step 4

The officer decides that he needs to add additional search criteria. Interpreting what the eyewitness told him, he types in '2000' for the *Year*, chooses a green color in the *Color* field, and clicks the *Search* button for a second time.

	State	License Plate Number	Make	Model	Year	Type	VIN	Owner's Zip Code	Color	R
		4A			2000	SUV				
	AR	W9PEINA	BMW	X5	2000	SUV	GUAT7IGD4AFXAFS4P	72342	Olive	S
	IL	0AE6P4A	Geo	Tracker	1996	SUV	P4R5HOV2WYPPEZXS4	62376	Olive	M
	CO	6MOX9QE	Isuzu	Trooper	2000	SUV	F452M7HDGGMZHNCYN	81227	Dark Gray	Te
	IA	YG5MHG2	Isuzu	Trooper	2000	SUV	530J6QASORNUQR5ST	52208	Dark Gray	G
	ND	4BJMCG	Isuzu	VehiCROSS	2001	SUV	OEB9F668BPSPV626W	58647	Olive	C
	FL	ULCF934	Mazda	Tribute	2001	SUV	BS24D8QSZ4J42P28D	34777	Olive	A M
	TX	CXVDYJ	Infiniti	QX4	2001	SUV	GEPGHVTQDXFM5X7DB	76953	Olive	D
	SD	PK1GPGZ	Mitsubishi	Montero Sport	2000	SUV	ESVPZO098C40D856X	57528	Brown	Lo
	PA	H331GUD	Chevrolet	Blazer	2000	SUV	KIKFM9NQMPSP50KRW	18469	Light Brown	C
	NV	MMAD4A	Chevrolet	Tracker	2003	SUV	X9MFGOND1LFSPZ5JO	89502	Dark Brown	A
	VT	FA5YUIT	Honda	CR-V	2001	SUV	S813LEERFBUYIX262	5669	Dark Gray	A
	ME	LH17PK	Toyota	Sequoia	2002	SUV	8EIO5TH2L5I5V4K6H	4903	Olive	A
	AZ	ZQCQQDC	GMC	Jimmy	2000	SUV	55GOER5GSRNEL940P	85099	Teal	B
	NJ	CUGDEH	GMC	Jimmy	2000	SUV	YFRQS2PG13WG55JLZ	7417	Teal	S

Step 5

Looking at his new search results, the officer sees that there are several Olive colored SUVs with a year close to 2000. The first result has a *License Plate Number* ending in 'NA', which could have been misread by the eyewitness as '4A'. Since he's not sure on how good the information was from the eyewitness, the officer decides to change the confidence values of his search terms by clicking on each column name and moving the *Confidence* slider bar. He changes the confidence on the *License Plate Number* column to 75%, the confidence on *Year* to 25%, the confidence on *Type* to 90%, and the confidence on *Color* to 50%.

Confidence: 50 %

Year (25%)	Type (90%)	VIN	Owner's Zip Code	Color (50%)	Register
2000	SUV				
2000	SUV	GUAT7IGD4AFXAFS4P	72342	Olive	Setiawan
1996	SUV	P4R5HOV2WYPPEZXS4	62376	Olive	Mercade

Step 6

The officer clicks *Search* for a third time and observes the new results. He sees that the top match is an Olive 1996 Geo Tracker with a *License Plate* of 0AE6P4A. Marvel gave it a composite score of 69.4% due to a 100% match on *Type*, 84.5% match on *Color*, 78.9 % match on *Year*, and 19.3% match on *License Plate Number*.

	State	License Plate Number (75%)	Make	Model	Year (25%)	Type (90%)	VIN	Owner's Zip Code	Color (50%)
		4A			2000	SUV			
69.4%	IL	0AE6P4A	Geo	Tracker	1996	SUV	P4R5HOV2WYPPEZXS4	62376	Olive
			Chevrolet	Tracker	2003	SUV	X9MFGOND1LFSPZ5JO	89502	Dark Brown
			BMW	X5	2000	SUV	GUAT7IGD4AFXAFS4P	72342	Olive
			Isuzu	VehiCROSS	2001	SUV	OEB9F668BPSPV626W	58647	Olive
			Mazda	Tribute	2001	SUV	BS24D8QSZ4J42P28D	34777	Olive
	TX	CXVDYJ	Infiniti	QX4	2001	SUV	GEPGHVTQDXFM5X7DB	76953	Olive
	CO	6MOX9QE	Isuzu	Trooper	2000	SUV	F452M7HDGGMZHNCYN	81227	Dark Gray
	IA	YG5MHG2	Isuzu	Trooper	2000	SUV	530J6QASORNUQR5ST	52208	Dark Gray
	ME	LH17PK	Toyota	Sequoia	2002	SUV	8EIO5TH2L5I5V4K6H	4903	Olive
	VT	FA5YUIT	Honda	CR-V	2001	SUV	S813LEERFBUYIX262	5669	Dark Gray
	SD	PK1GPGZ	Mitsubishi	Montero Sport	2000	SUV	ESVPZO098C40D856X	57528	Brown

Type	100.0%
Color	84.5%
Year	78.9%
License Plate Number	19.3%

Step 7

The officer takes down the vehicle VIN number and proceeds to track down its registered owner.

Hardware and Software Requirements

The demonstration will require 3-4 laptop computers which will be provided by the authors/demonstrators of this proposal. The demonstration will require the IAAI-08 to provide **No** hardware or software in support of this demonstration.