1. Research

1.1. Web Crawler

Research has been carried out into improving automatic web crawling. An analysis showed a number of data extraction problems with existing techniques for the kinds of websites relevant to SCIIMS (including online classified advertisements, social networks and job websites). An improved algorithm has been produced with beyond state of the art features. Subsequent experiments have shown that this achieves a success rate in excess of 85%.

Research has also been carried into improved automated Web browsing. Most previous web automation systems use conventional browsers to automate navigation sequences. This is computationally expensive in terms of both CPU and memory, especially with the new breed of websites (Ajax, etc.). This causes scalability problems where many browsers may be executing at the same time, which is the typical case in web automation applications. The approach is based on the insight that, in the case of automated navigation, browsing sequences are known in advance. Therefore it is possible to carry out a test execution to find what elements in the website pages are needed to execute the target navigation sequence and then only load and execute the required elements for subsequent executions. This has proved to be very effective, for instance automated web navigations with Microsoft Explorer are over 5 times slower than with the specialised browser. Figure 1 shows an example of this: to correctly process a click on the greyed A node, only the greyed nodes are needed (detecting the required nodes involves a complex process described in other documents). The other nodes/scripts need not be loaded or executed, resulting in a significant decrease of memory and CPU usage.

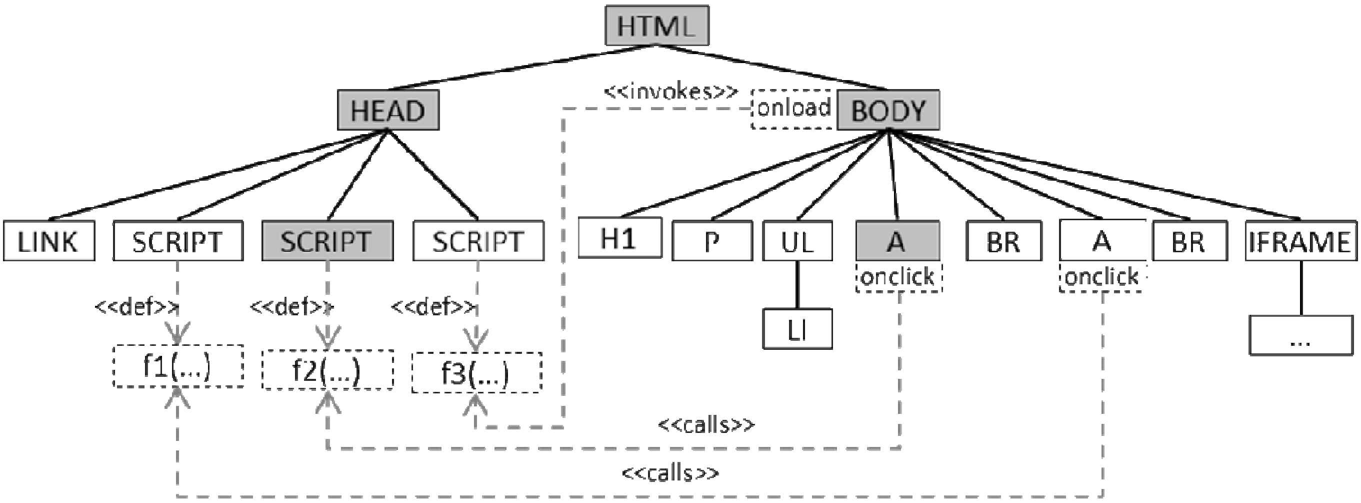


Figure 1 SCIIMS Web Navigation

1.2. Data Mining

Research has been carried out into Entity Resolution (ER) which covers the problem of identifying distinct representations of real-world entities in heterogeneous databases. This examined the trade-off between the storage needs, performance, and the efficiency of Entity Resolution. Consequently several research papers have been published. An example of this is the paper Infrastructures and Bounds for Distributed Entity Resolution which includes Figure 2 showing the execution time for different ER methods and number of records.

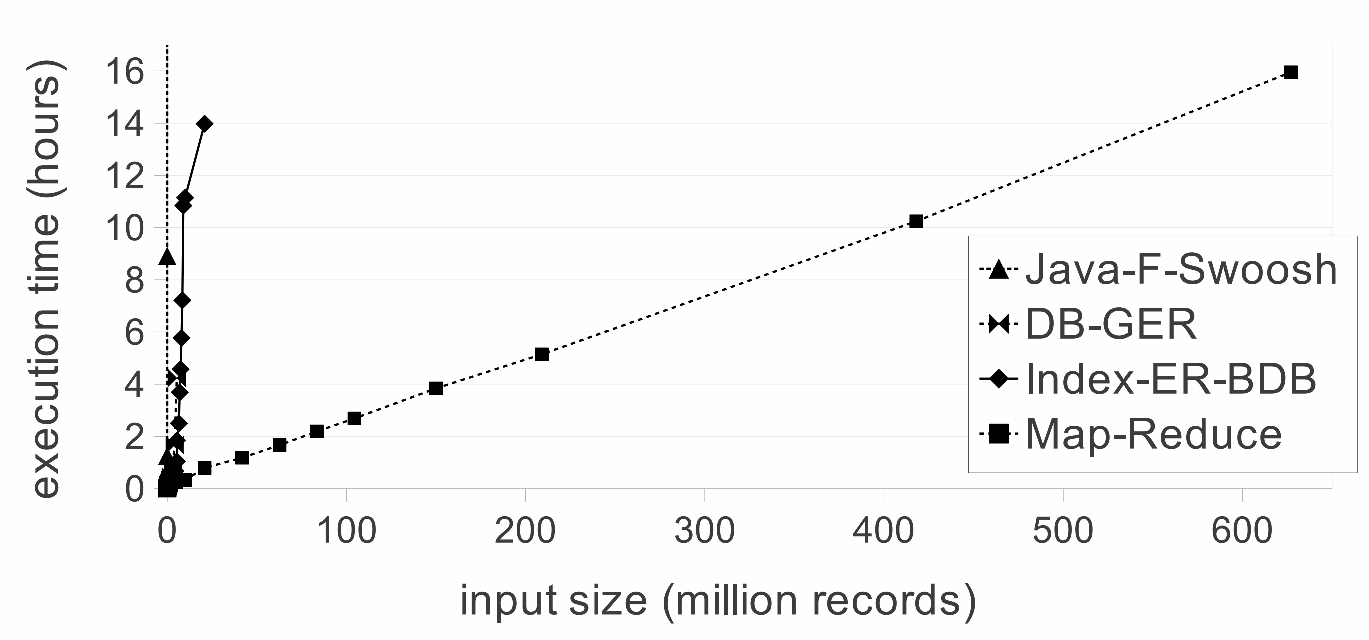
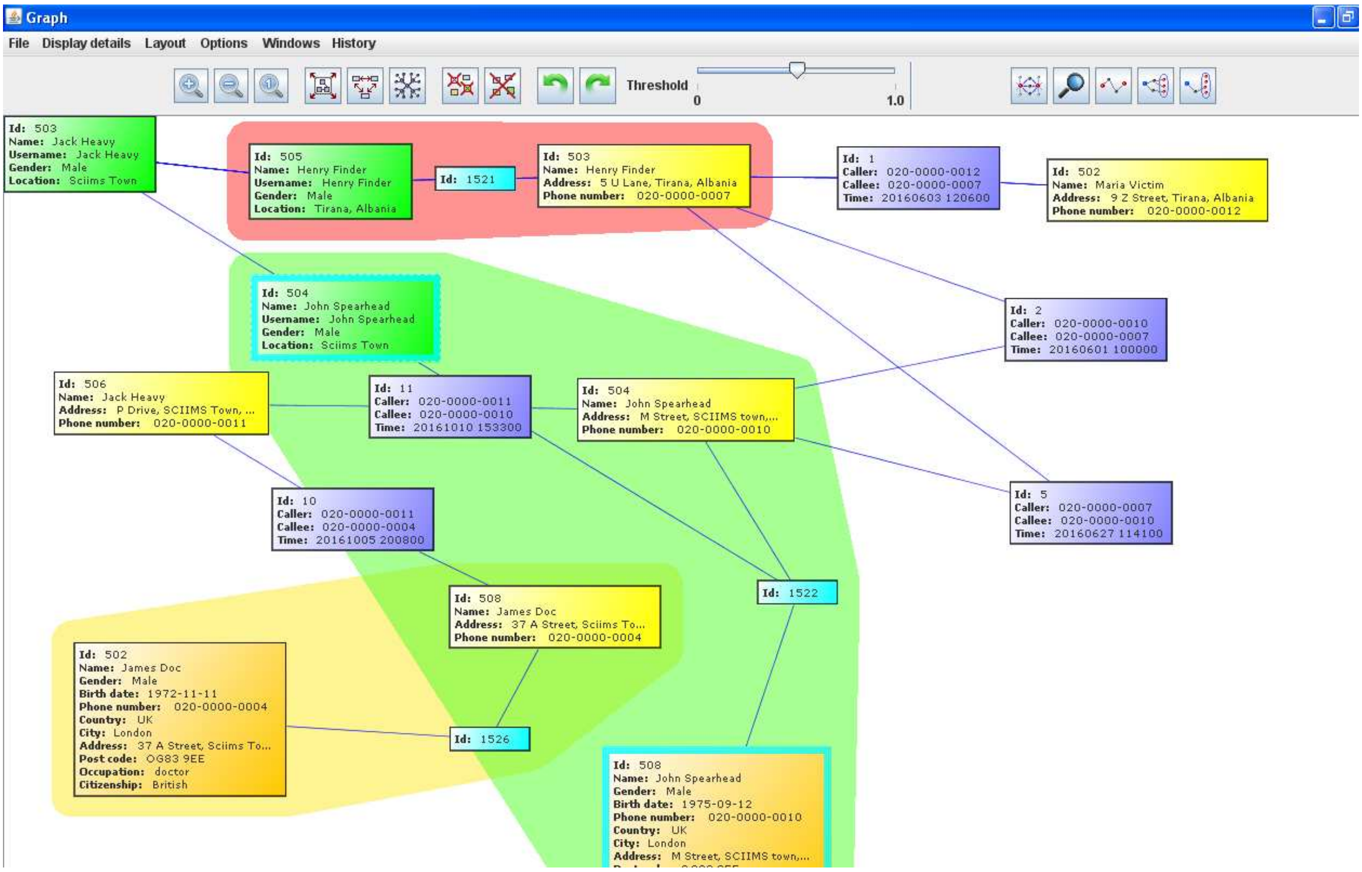


Figure 2 Entity Resolution Execution Time Against Size Entity Resolution

Research has been carried out into techniques for the visualisation of heterogeneous data sets as a network of entity nodes with arbitrary connections. As a result improved and new tools for visual analytics have been produced. An example of this from the Prototype/Demonstration System (described in Section 3) is shown in

Figure 3 .



Id: 503
Name: Jack Heavy
Username: Jack Heavy
Gender: Male
Location: Sciims Town

Id: 505
Name: Henry Finder
Username: Henry Finder
Gender: Male
Location: Tirana, Albania

Id: 1521

Id: 503
Name: Henry Finder
Address: 5 U Lane, Tirana, Albania
Phone number: 020-0000-0007

Id: 1
Caller: 020-0000-0012
Callee: 020-0000-0007
Time: 20160603 120600

Id: 502
Name: Maria Victim
Address: 9 Z Street, Tirana, Albania
Phone number: 020-0000-0012

Id: 504
Name: John Spearhead
Username: John Spearhead
Gender: Male
Location: Sciims Town

Id: 2
Caller: 020-0000-0010
Callee: 020-0000-0007
Time: 20160601 100000

Id: 506
Name: Jack Heavy
Address: P Drive, SCIIMS Town, ...
Phone number: 020-0000-0011

Id: 11
Caller: 020-0000-0011
Callee: 020-0000-0010
Time: 20161010 153300

Id: 504
Name: John Spearhead
Address: M Street, SCIIMS town,...
Phone number: 020-0000-0010

Id: 5
Caller: 020-0000-0007
Callee: 020-0000-0010
Time: 20160627 114100

Id: 10
Caller: 020-0000-0011
Callee: 020-0000-0004
Time: 20161005 200800

Id: 1522

Id: 508
Name: James Doc
Address: 37 A Street, Sciims To...
Phone number: 020-0000-0004

Id: 502
Name: James Doc
Gender: Male
Birth date: 1972-11-11
Phone number: 020-0000-0004
Country: UK
City: London
Address: 37 A Street, Sciims To...
Post code: OG83 9EE
Occupation: doctor
Citizenship: British

Id: 1526

Id: 508
Name: John Spearhead
Gender: Male
Birth date: 1975-09-12
Phone number: 020-0000-0010
Country: UK
City: London
Address: M Street, SCIIMS town,...

Figure 3 Graphical View Example

2. Dissemination

2.1. Newsletter and Poster

A newsletter and poster have been produced alongside the web site (www.sciims.eu) to communicate the work carried out on the project to interested parties (see Figure 4).

SCIIMS - ISSUE 2 - AUGUST 2011

SCIIMS

About SCIIMS
Turn to page 2
for full details

Consortium
Achievements to Date
Turn to page 6
for full details

INTRODUCTION

Welcome to the SCIIMS Newsletter which has been produced in response to your request for more information. SCIIMS has been operating since November 2009 and has made significant progress during this time. Some of the achievements include producing operational and adversary models, a set of user and system requirements and carrying out novel and beyond state of the art research. Currently the focus is on the system design and the ontology that underpins SCIIMS. This issue describes the achievements of the consortium since the start of the project and contains some brief articles about some of the latest research.

Latest News Headlines
Summary

SCIIMS has finalised its hardware and software architecture

SCIIMS conducts research into entity resolution and visualisation techniques

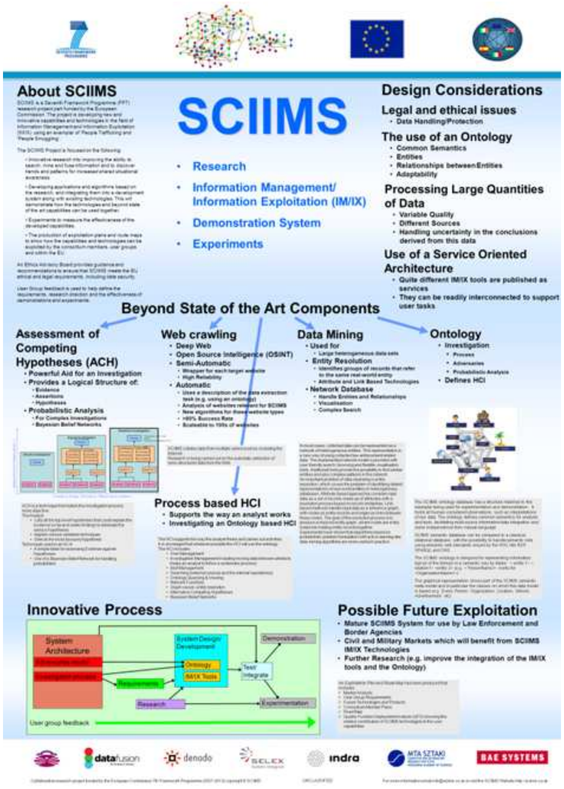
SCIIMS investigates novel data visualisation techniques

SCIIMS uses an Ontology

SCIIMS investigates novel web data extraction algorithms

SCIIMS starts testing novel web data extraction techniques

In this issue
Introduction Page 1
Latest News Headlines Summary Page 1
About SCIIMS Page 2
Latest News Page 2
EU Seventh Framework Programme Page 2
Latest Research Page 4
Consortium Achievements Page 6
Individual Consortium Achievements Page 6
Events/exhibitions/conferences Page 7
Collaborating organisation Page 8



About SCIIMS

SCIIMS

- Research
- Information Management/ Information Exploitation (IM/IX)
- Demonstration System
- Experiments

Design Considerations

Legal and ethical issues
- Data Handling/Protection

The use of an Ontology
- Common Semantics
- Entities
- Relationships between Entities
- Adaptability

Processing Large Quantities of Data
- Variable Quality
- Different Sources
- Handling uncertainty in the conclusions derived from this data

Use of a Service Oriented Architecture
- Quite different IM/IX tools are published as services
- They can be readily interconnected to support user tasks

Beyond State of the Art Components

Assessment of Competing Hypotheses (ACH)
- Powerful Aid for an Investigation
- Provides a Logical Structure of:
  - Evidence
  - Assertions
  - Hypotheses
- Probabilistic Analysis
  - For Complex Investigations
  - Bayesian Belief Networks

Web crawling
- Deep Web
- Open Source Intelligence (OSINT)
- Semi-Automatic
  - Wrapper for each target website
  - High Reliability
- Automatic
  - Uses a description of the data extraction task (e.g. using an ontology)
  - Analysis of websites relevant for SCIIMS
  - New algorithms for these website types
  - >90% Success Rate
  - Scaleable to 100s of websites

Data Mining
- Used for
  - Large heterogeneous data sets
- Entity Resolution
  - Identifies groups of records that refer to the same real-world entity
  - Attribute and Link Based Technologies
- Network Database
  - Handle Entities and Relationships
  - Visualisation
  - Complex Search

Ontology
- Investigation
  - Process
  - Adversaries
  - Probabilistic Analysis
- Defines HCI

Process based HCI
- Supports the way an analyst works
- Investigating an Ontology based HCI

Innovative Process

Possible Future Exploitation
- Mature SCIIMS System for use by Law Enforcement and Border Agencies
- Civil and Military Markets which will benefit from SCIIMS IM/IX Technologies
- Further Research (e.g. improve the integration of the IM/IX tools and the Ontology)

Figure 4 Newsletter and Poster

3. Prototype/Demonstration System

3.1. Modules and Basic Architecture

A working Prototype/Demonstration System has been produced to demonstrate the capabilities of SCIIMS and also for system level experiments. Figure 5 shows the modules and basic architecture of the Prototype/Demonstration system.

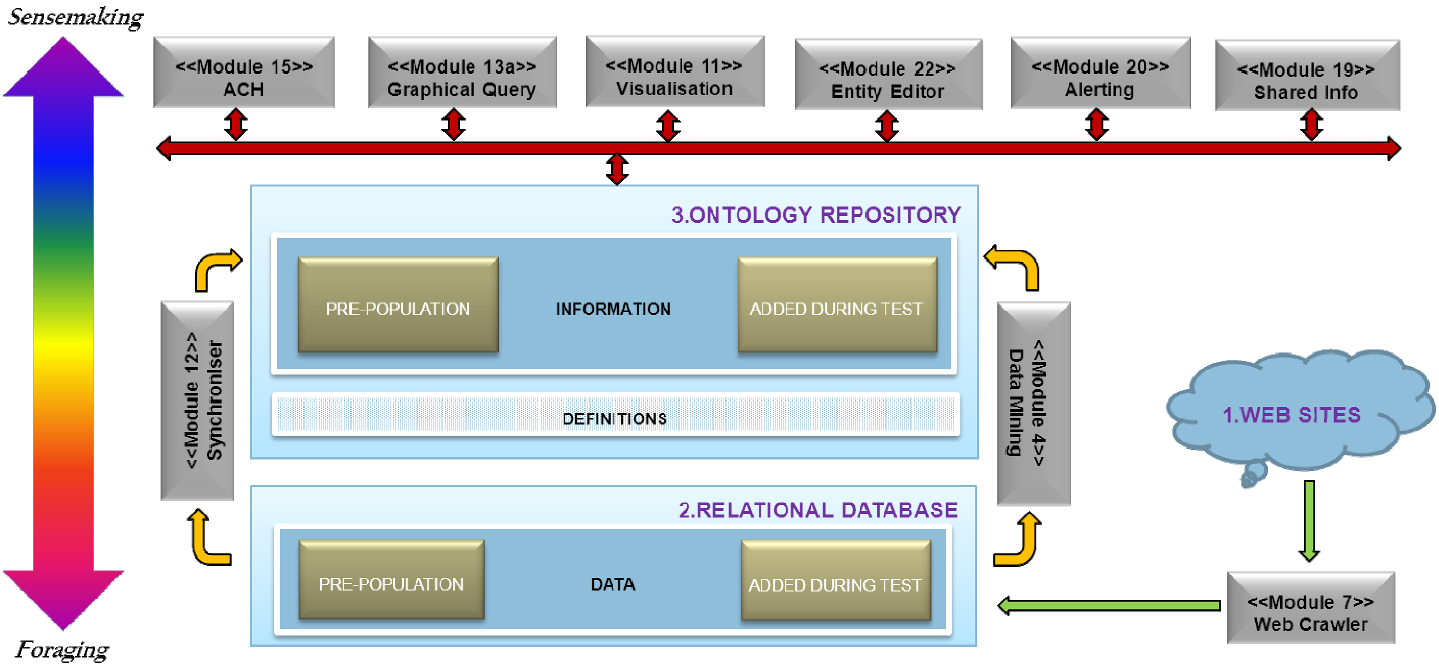


Figure 5 SCIIMS Demonstration System Components

Supporting this are the:

a) Data Services Layer (Module 5) which combines, integrates and transforms information obtained from the different data sources for use by the SCIIMS modules. This includes data virtualization technology for RDF / OWL which is beyond state of the art,

b) Integration Module (Module 17) which deals with the asynchronous aspects (for instance Data Mining and Web Crawling). To support the Service Orientated Architecture (SOA) an Enterprise Service Bus (ESB) is used.

Not shown on this diagram is the Netica tool which is used to construct graphically Bayesian Belief networks as necessary to support an investigation (see

Figure 6)

3.2. Organising a Criminal Investigation

A number of techniques have been identified for providing computer-assistance to an Investigator and combined into a conceptual system of thinking.

The originating techniques are:

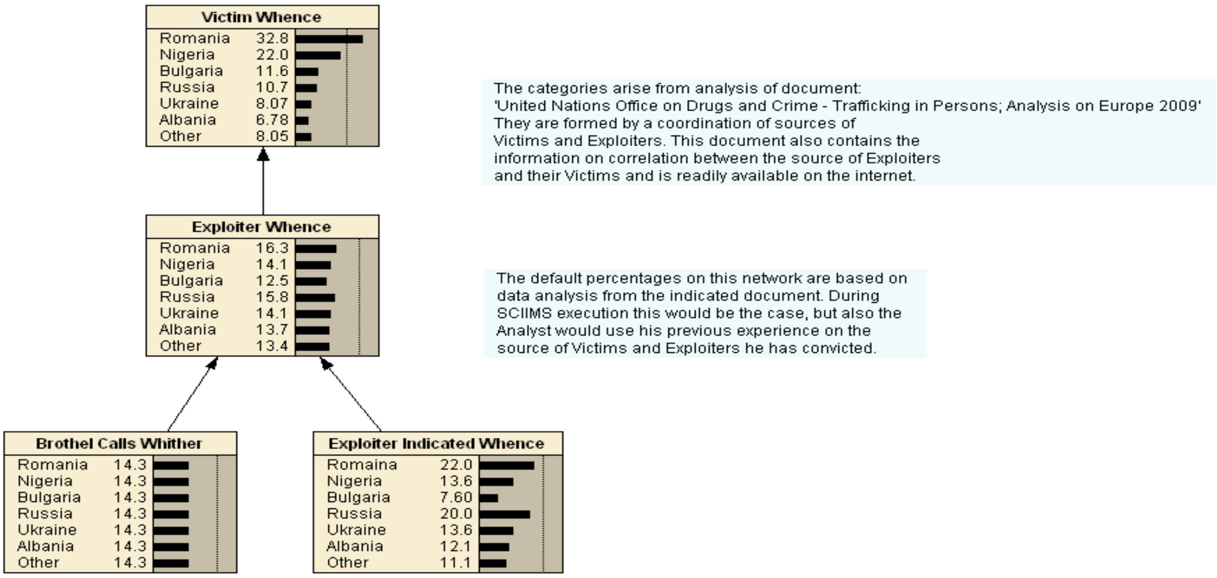
a) **Pirolli-Card** (P-C) – a cognitive model that shows how an investigator typically conducts an Investigation

b) **Goal Structured Notation** (GSN) – a general information-representation technique that can be used in SCIIMS to show a hierarchy of findings, linking evidence to actionable conclusions

c) **Assessment of Competing Hypotheses** (ACH) – an investigative technique that helps to avoid making premature conclusions. The Investigator is supported in maintaining several hypotheses that explain some finding, only excluding those explanations when evidence contradicts them.

Bayesian Belief Networks **(BBN) – a probabilistic technique used to reason with a system of related hypotheses. In SCIIMS it is used to process the hypotheses of an ACH when the ACH would otherwise be too difficult to consider mentally (see**

d) Figure 6)



The categories arise from analysis of document:
'United Nations Office on Drugs and Crime - Trafficking in Persons; Analysis on Europe 2009'
They are formed by a coordination of sources of
Victims and Exploiters. This document also contains the
information on correlation between the source of Exploiters
and their Victims and is readily available on the internet.

The default percentages on this network are based on
data analysis from the indicated document. During
SCIIMS execution this would be the case, but also the
Analyst would use his previous experience on the
source of Victims and Exploiters he has convicted.

Figure 6 Example BBN

SCIIMS allows a simple investigation to be shown and then enhanced using GSN, ACH and BBN techniques as the need arises. This is depicted in

Figure 7 below.

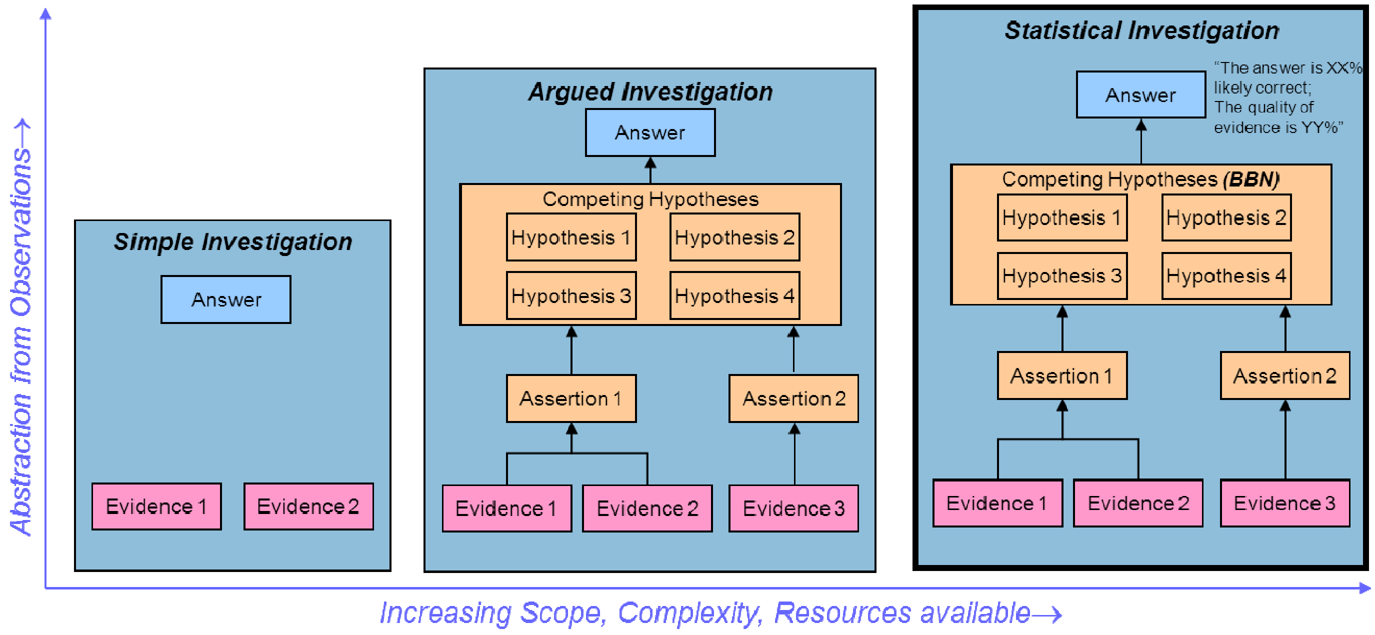


Figure 7 Increasingly Complex Investigations

Figure 8 shows how an argument can be represented on the SCIIMS Prototype/Demonstration System using GSN.

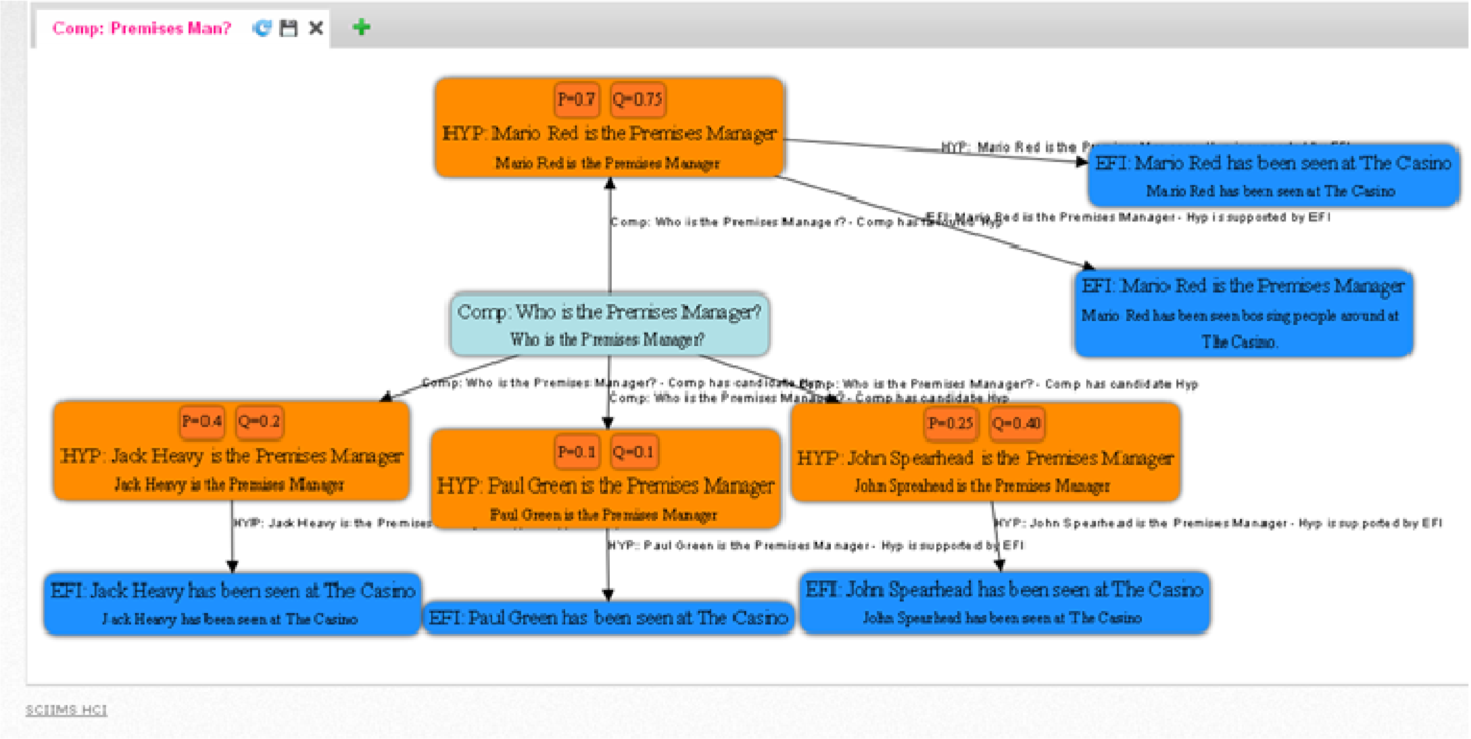

Comp: Premises Man?
P=0.7
Q=0.75
HYP: Mario Red is the Premises Manager
Mario Red is the Premises Manager
HYP: Mario Red is the P
EFI: Mario Red has been seen at The Casino
Mario Red has been seen at The Casino
Comp: Who is the Premises Manager? - Comp has candidate Hyp
EFI: Mario Red is the Premises Manager - Hyp is supported by EFI
EFI: Mario Red is the Premises Manager
Mario Red has been seen bossing people around at The Casino.
Comp: Who is the Premises Manager?
Who is the Premises Manager?
Comp: Who is the Premises Manager? - Comp has candidate Hyp
Comp: Who is the Premises Manager? - Comp has candidate Hyp
Comp: Who is the Premises Manager? - Comp has candidate Hyp
P=0.4
Q=0.2
HYP: Jack Heavy is the Premises Manager
Jack Heavy is the Premises Manager
P=0.1
Q=0.1
HYP: Paul Green is the Premises Manager
Paul Green is the Premises Manager
P=0.25
Q=0.40
HYP: John Spearhead is the Premises Manager
John Spreahead is the Premises Manager
HYP: Jack Heavy is the Premises
HYP: Paul Green is the Premises Manager - Hyp is supported by EFI
HYP: John Spearhead is the Premises Manager - Hyp is supported by EFI
EFI: Jack Heavy has been seen at The Casino
Jack Heavy has been seen at The Casino
EFI: Paul Green has been seen at The Casino
EFI: John Spearhead has been seen at The Casino
John Spearhead has been seen at The Casino
SCIIMS HCI

Figure 8 Example of an Analyst using the ACH Investigation Structure in the Ontology

3.3. Ontology

Central to the SCIIMS Prototype/demonstration System is the Ontology which has the following major sections (see

Figure 9):

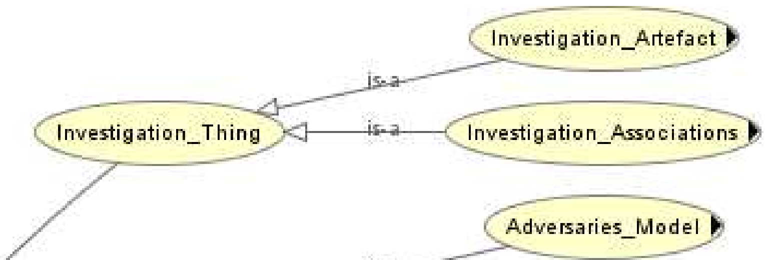
a) Operational - aligned with the Adversaries Model and containing persons, places, relationships, events etc.

b) Process – for the Pirolli and Card business process, user identification etc.

c) Investigation – for shoebox data, searches, alerts etc.

d) HCI

e) BBN



.

Figure 9 SCIIMS Ontology

Figure 10 illustrates how an investigation is structured within the Ontology both for entities for hypotheses, evidence file items, shoebox items etc. and their associations.

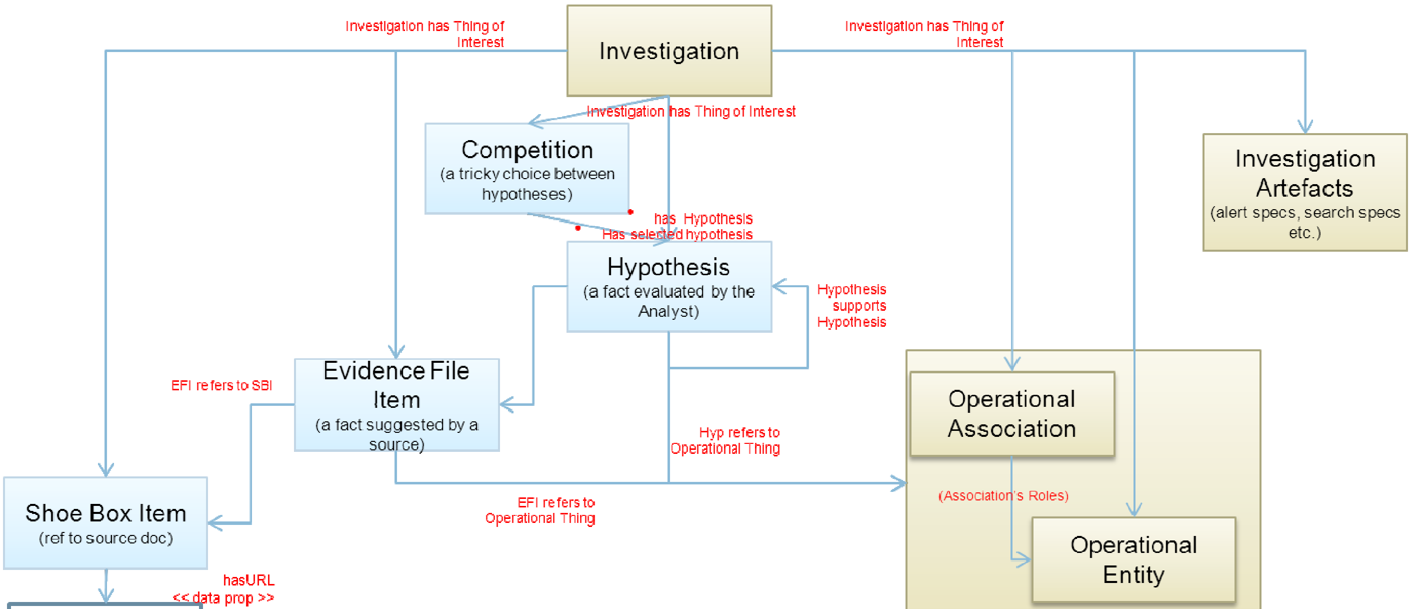
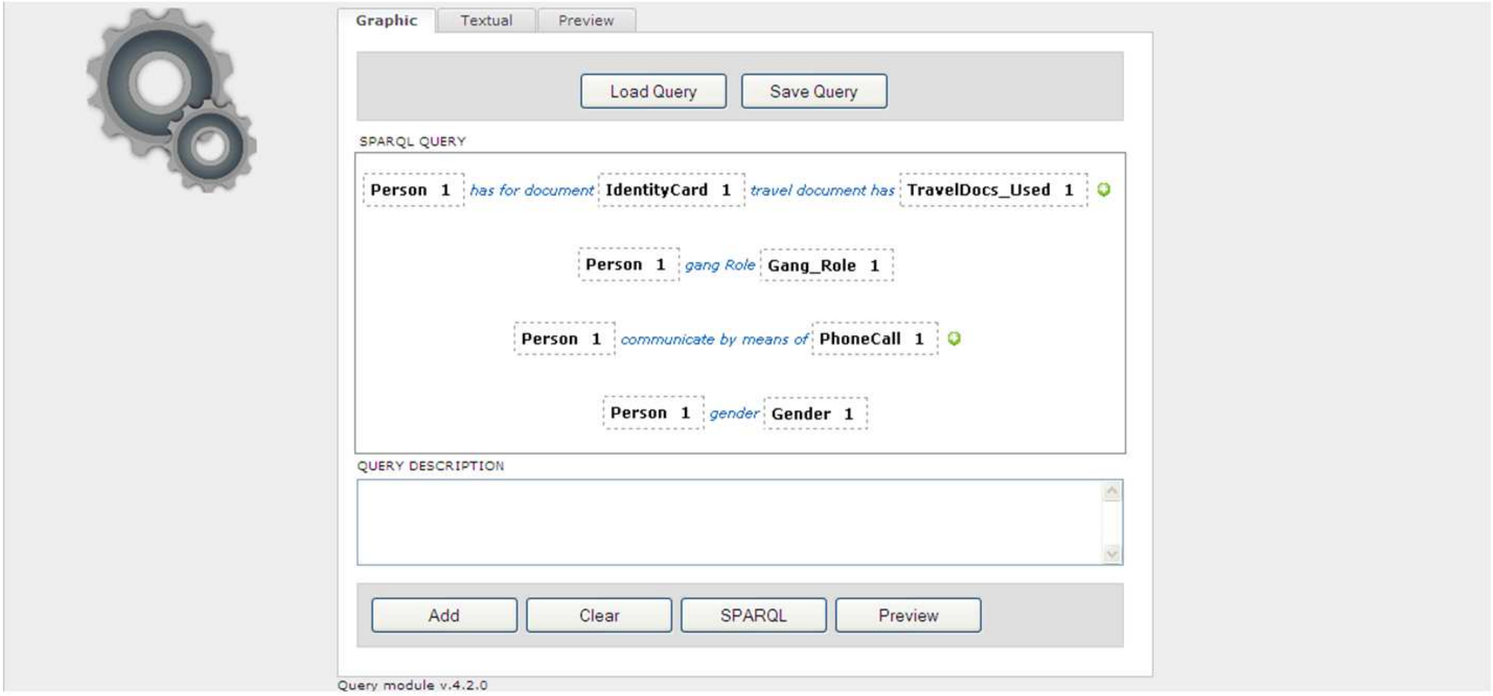

Investigation
Investigation has Thing of Interest
Investigation has Thing of Interest
Investigation has Thing of Interest
Competition
(a tricky choice between hypotheses)
Investigation Artefacts
(alert specs, search specs etc.)
has Hypothesis
Has selected hypothesis
Hypothesis
(a fact evaluated by the Analyst)
Hypothesis supports Hypothesis
EFI refers to SBI
Evidence File Item
(a fact suggested by a source)
Hyp refers to Operational Thing
Operational Association
(Association's Roles)
EFI refers to Operational Thing
Shoe Box Item
(ref to source doc)
Operational Entity
hasURL
<< data prop >>

Figure 10 Investigation Structure in the Ontology

Associated with the Ontology is the Graphical Query Search Module. This can be used to construct detailed searches both graphically and using SPARQL for information available from the Ontology. Figure 11 shows an example of the module and the results.

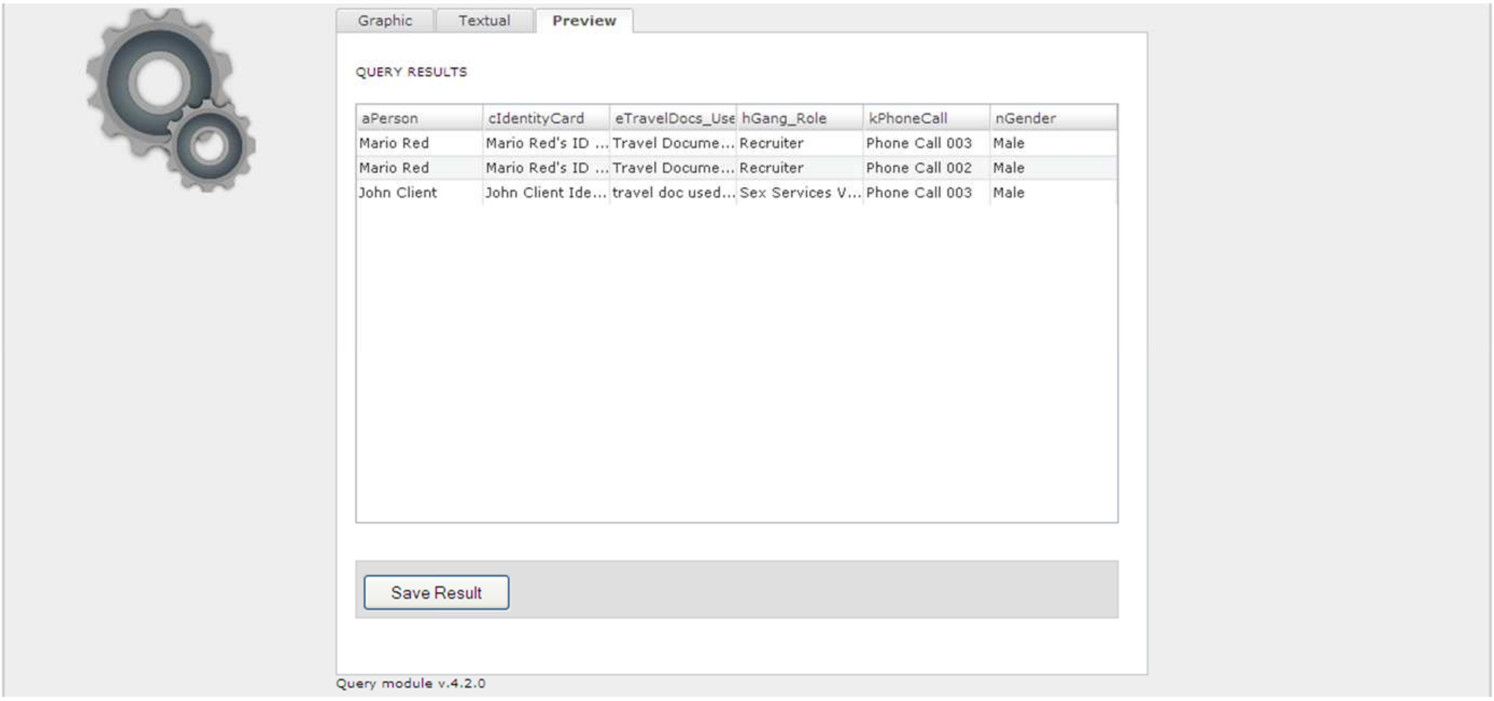


Person 1 *has for document* IdentityCard 1 *travel document has* TravelDocs_Used 1

Person 1 *gang Role* Gang_Role 1

Person 1 *communicate by means of* PhoneCall 1

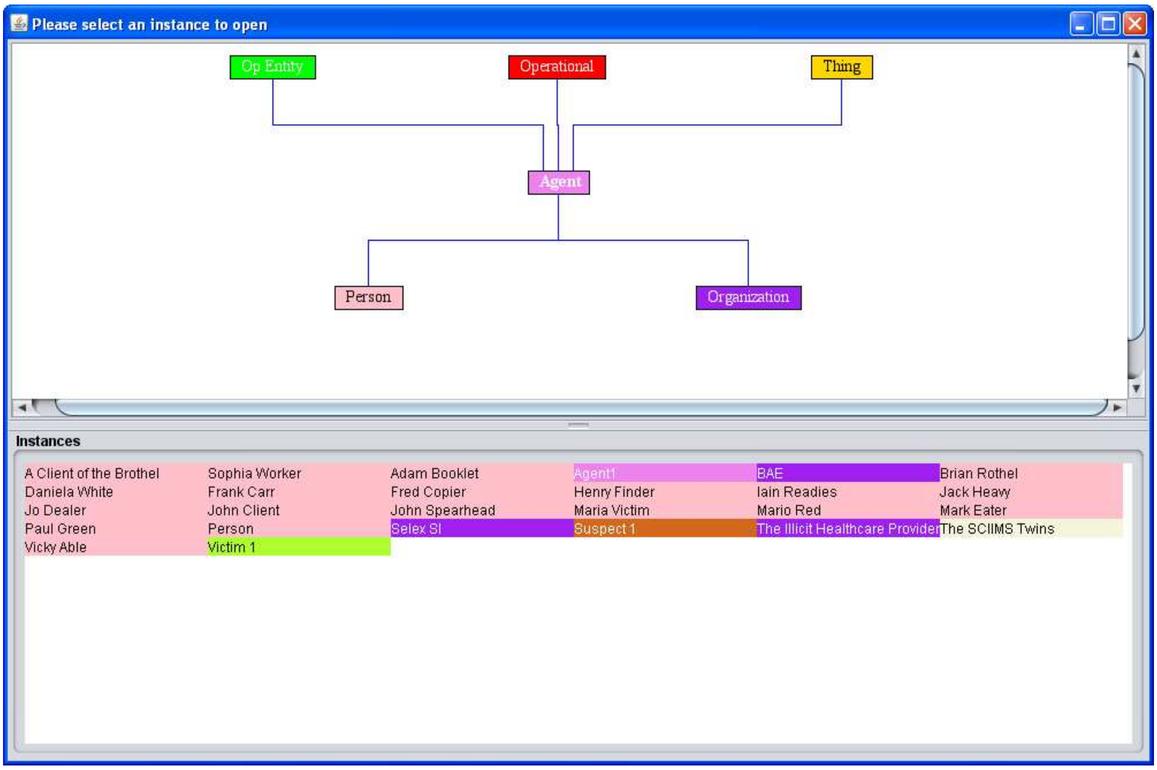
Person 1 *gender* Gender 1



aPerson	cIdentityCard	eTravelDocs_Use	hGang_Role	kPhoneCall	nGender
Mario Red	Mario Red's ID ...	Travel Docume...	Recruiter	Phone Call 003	Male
Mario Red	Mario Red's ID ...	Travel Docume...	Recruiter	Phone Call 002	Male
John Client	John Client Ide...	travel doc used...	Sex Services V...	Phone Call 003	Male

Figure 11 SCIIMS Graphical query Module

The Entity Editor allows a user to navigate an Ontology (a graphical structure is displayed - see Figure 12) and provides direct read and write access.



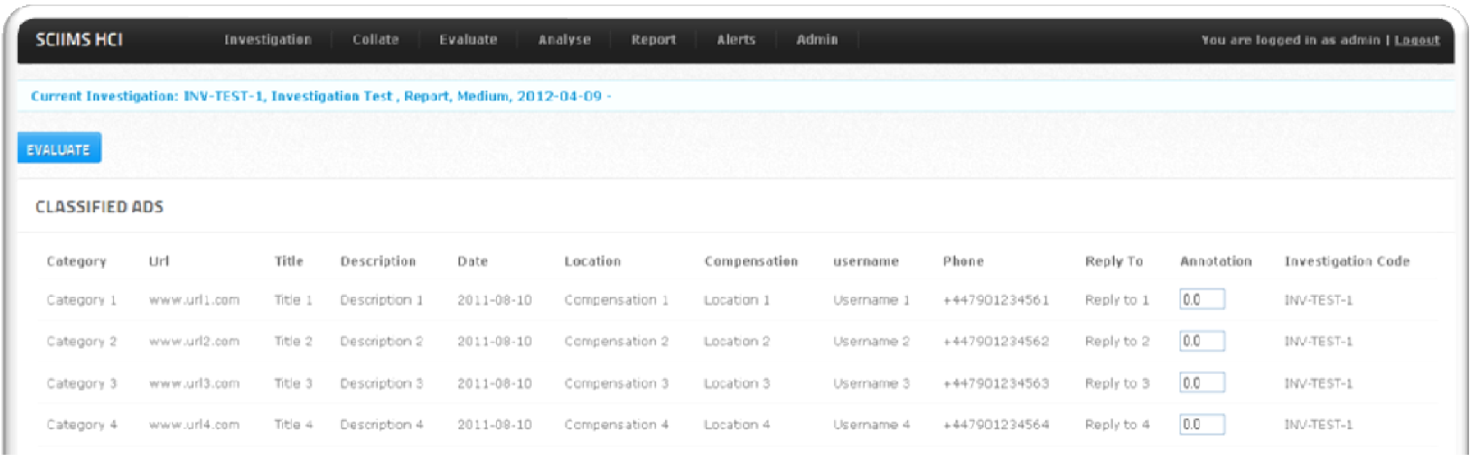
A Client of the Brothel	Sophia Worker	Adam Booklet	Agent1	BAE	Brian Rothel
Daniela White	Frank Carr	Fred Copier	Henry Finder	Iain Readies	Jack Heavy
Jo Dealer	John Client	John Spearhead	Maria Victim	Mario Red	Mark Eater
Paul Green	Person	Selex SI	Suspect 1	The Illicit Healthcare Provider	The SCIIMS Twins
Vicky Able	Victim 1				

Figure 12 Entity Editor

.

3.4. HCI

Figure 13 shows an example of the SCIIMS Prototype/Demonstration System HCI, in this case the HCI for the classification of advertisements.



Current Investigation: INV-TEST-1, Investigation Test , Report, Medium, 2012-04-09 -

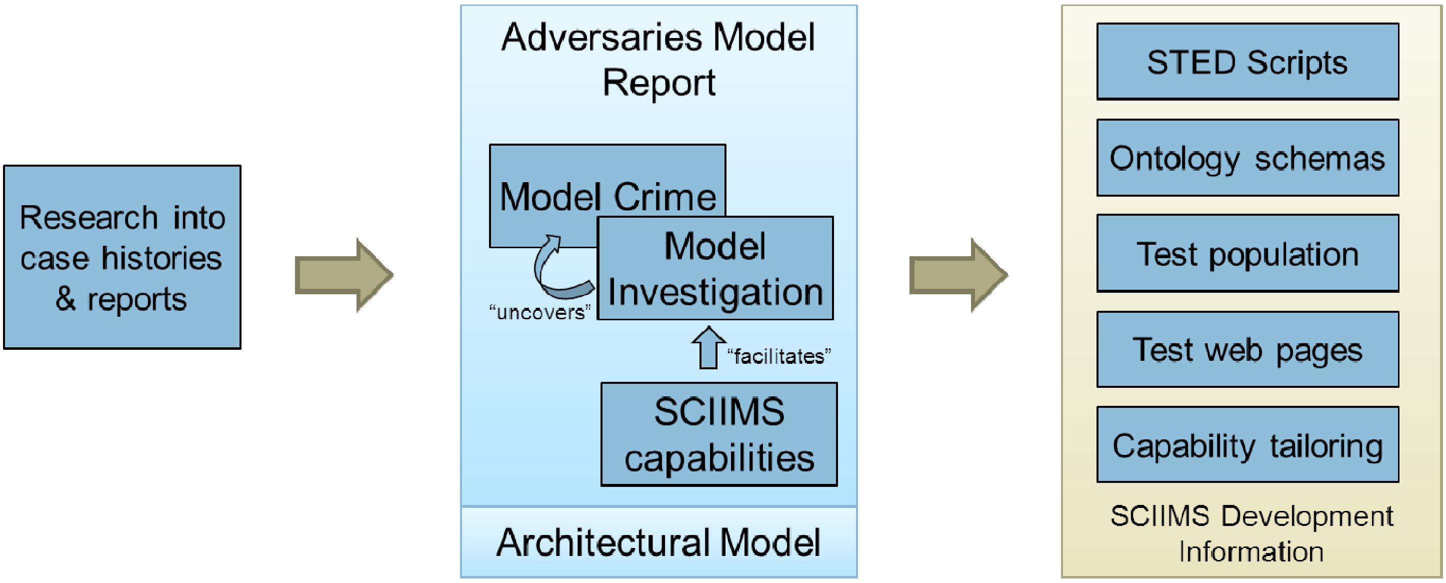
CLASSIFIED ADS

Category	Url	Title	Description	Date	Location	Compensation	username	Phone	Reply To	Annotation	Investigation Code
Category 1	www.url1.com	Title 1	Description 1	2011-08-10	Compensation 1	Location 1	Username 1	+447901234561	Reply to 1	0.0	INV-TEST-1
Category 2	www.url2.com	Title 2	Description 2	2011-08-10	Compensation 2	Location 2	Username 2	+447901234562	Reply to 2	0.0	INV-TEST-1
Category 3	www.url3.com	Title 3	Description 3	2011-08-10	Compensation 3	Location 3	Username 3	+447901234563	Reply to 3	0.0	INV-TEST-1
Category 4	www.url4.com	Title 4	Description 4	2011-08-10	Compensation 4	Location 4	Username 4	+447901234564	Reply to 4	0.0	INV-TEST-1

Figure 13 Example Prototype/Demonstration System HCI Screen

4. System Test, Demonstrations and Experiments (STED)

A number of demonstrations and system level experiments (where “human in the loop” experiments are carried out).using the Prototype/Demonstration System have been conducted to gain feedback on the various SCIIMS capabilities The overall concept was to produce a test script based on the Adversaries Model Scenario (this describes in detail an example of an investigation into people trafficking) and use it as a basis for the System Tests, Demonstrations (which uses a subset of the tests) and experimentation Figure 14 illustrates how the Adversaries Model is used as a basis for the system test, demonstration and experimentation scripts as well as defining the test data for the Prototype/Demonstration System relational databases, ontology and web pages.



Research into case histories & reports

Adversaries Model Report

Model Crime

"uncovers"

Model Investigation

"facilitates"

SCIIMS capabilities

Architectural Model

STED Scripts

Ontology schemas

Test population

Test web pages

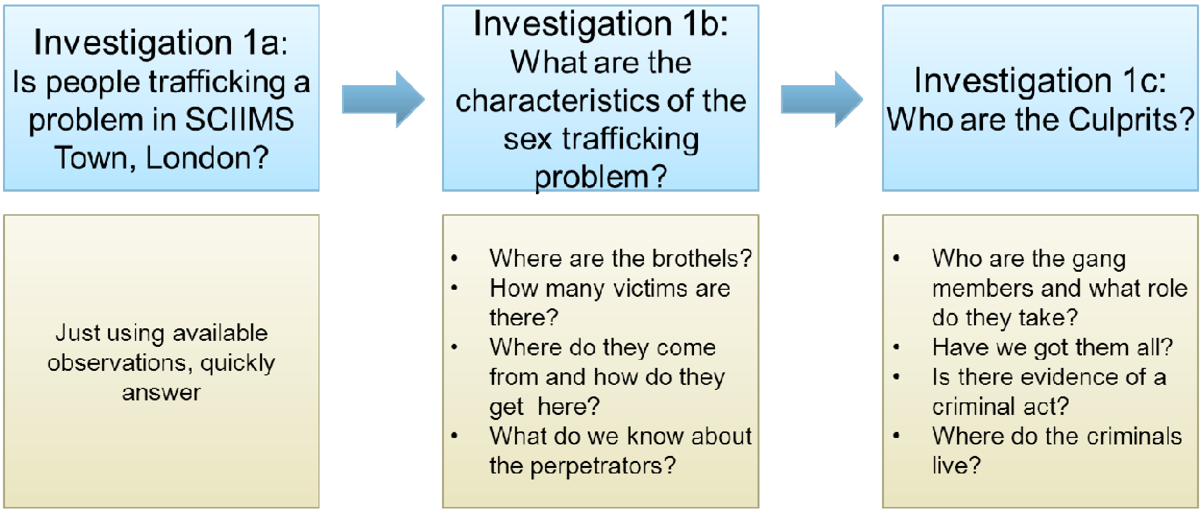
Capability tailoring

SCIIMS Development Information

- System test, experimentation and demonstration coordinated through architecture
- Basis for future rapid development process

Figure 14 Innovative Process for Developing SCIIMS

Figure 15 summarises the three stages to the Adversaries Model investigation. The separate test scripts were written to cover phases 1a, 1b and 1c which separate the scripts covering the foraging parts (e.g. web crawling and data mining) and the sense making scripts.



Investigation 1a: Is people trafficking a problem in SCIIMS Town, London?

Just using available observations, quickly answer

Investigation 1b: What are the characteristics of the sex trafficking problem?

- Where are the brothels?
- How many victims are there?
- Where do they come from and how do they get here?
- What do we know about the perpetrators?

Investigation 1c: Who are the Culprits?

- Who are the gang members and what role do they take?
- Have we got them all?
- Is there evidence of a criminal act?
- Where do the criminals live?

- *This sequence occurs over several weeks in August 2016 as the criminal details emerge*

Figure 15 Adversaries Model Investigation Scenario

4.1. Demonstration Results

Figure 16 shows an example of the results from the demonstrations. Anything above 0 is a useful capability.

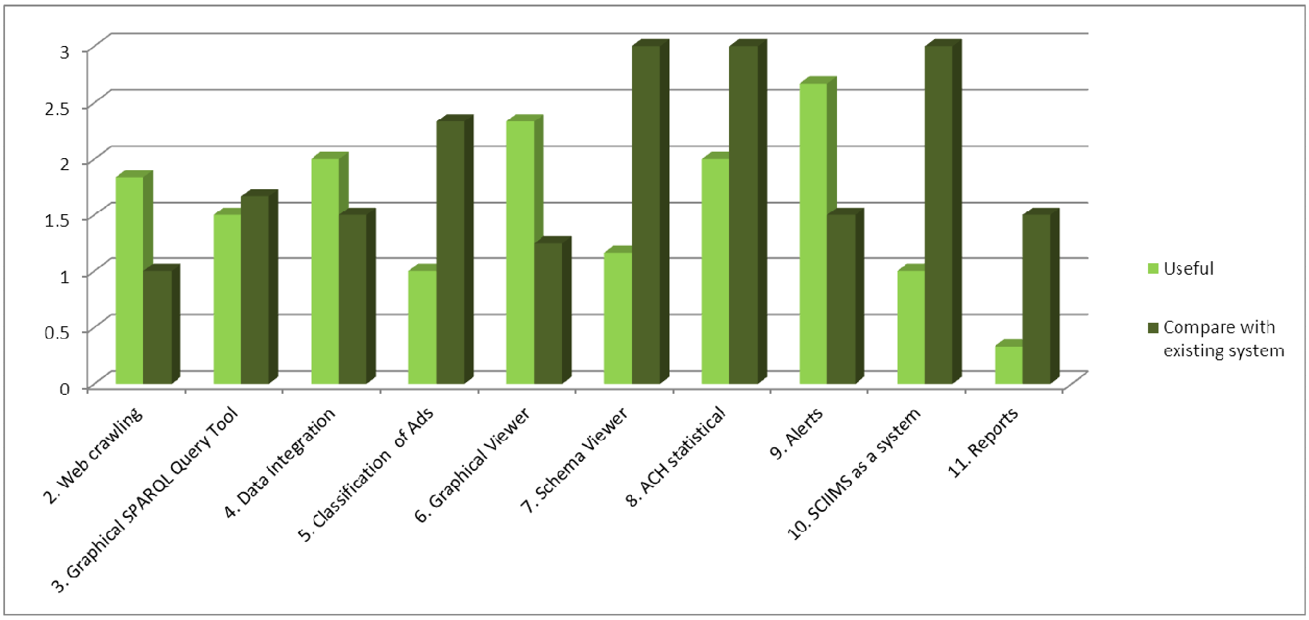


Figure 16 Usefulness and comparison with existing system

4.2. Experimentation Results

Figure 17 shows an example of the results from the experiments. Anything at or above the green line is easy to use.

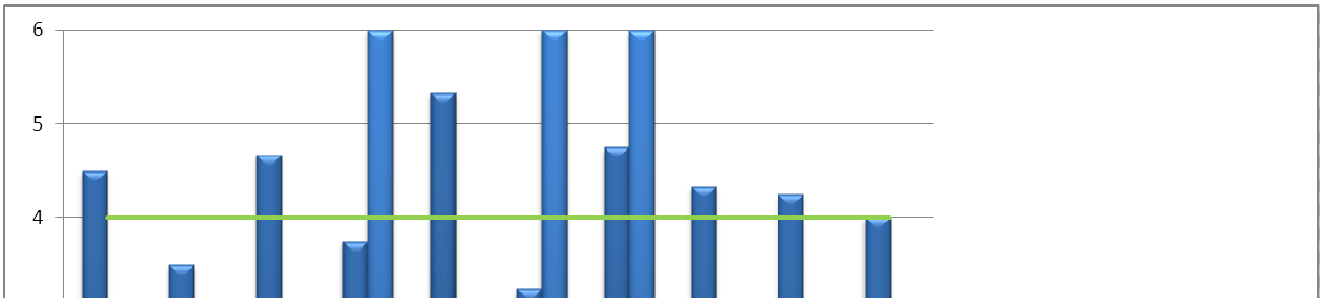


Figure 17 Ease of Use and Comparison with Existing Systems