Intelligent Information System Supporting
Observation, Searching and Detection for
Security of Citizens in Urban Environment

European Seventh Framework Programme
FP7-218086-Collaborative Project

XML Data Corpus: Report on methodology for collection, cleaning and unified representation of large textual data from various sources: news reports, weblogs, chat. WP4. D.4.1

The INDECT Consortium

AGH – University of Science and Technology, AGH, Poland
Gdansk University of Technology, GUT, Poland
InnoTec DATA GmbH & Co. KG, INNOTEC, Germany
IP Grenoble (Ensimag), INP, France
MSWiA[1] - General Headquarters of Police (Polish Police), GHP, Poland
Moviquity, MOVIQUITY, Spain
Products and Systems of Information Technology, PSI, Germany
Police Service of Northern Ireland, PSNI, United Kingdom
Poznan University of Technology, PUT, Poland
Universidad Carlos III de Madrid, UC3M, Spain
Technical University of Sofia, TU-SOFIA, Bulgaria
University of Wuppertal, BUW, Germany
University of York, UoY, Great Britain
Technical University of Ostrava, VSB, Czech Republic
Technical University of Kosice, TUKE, Slovakia
X-Art Pro Division G.m.b.H., X-art, Austria
Fachhochschule Technikum Wien, FHTW, Austria

[1] MSWiA (Ministerstwo Spraw Wewnętrznych i Administracji) – Ministry of Interior Affairs and Administration. Polish Police is dependent on the Ministry

Document Information

Contract Number	*218086*
Deliverable name	XML Data Corpus: Report on methodology for collection, cleaning and unified representation of large textual data from various sources: news reports, weblogs, chat.
Deliverable number	*D.4.1 (WP4)*
Editor(s)	Suresh Manandhar, University of York, suresh@cs.york.ac.uk
Author(s)	Ioannis Klapaftis, University of York giannis@cs.york.ac.uk Suresh Manandhar, University of York, suresh@cs.york.ac.uk Shailesh Pandey, University of York shailesh@cs.york.ac.uk
Reviewer(s)	Alan Frisch, University of York frisch@cs.york.ac.uk
Dissemination level	*CONFIDENTIAL*
Contractual date of delivery	*30/06/2009*
Delivery date	*30/06/2009*
Status	*final deliverable*
Keywords	XML Data Corpus, methodology for collection, news reports, weblogs, chat.

This project is funded under 7th Framework Program

Table of Contents

(This page is left blank intentionally)

1 Executive Summary

Security is becoming a weak point of energy and communications infrastructures, commercial stores, conference centers, airports and sites with high person traffic in general. Practically any crowded place is vulnerable, and the risks should be controlled and minimized as much as possible. Access control and rapid response to potential dangers are properties that every security system for such environments should have. The INDECT project is aiming to develop new tools and techniques that will help the potential end users in improving their methods for crime detection and prevention thereby offering more security to the citizens of the European Union.

In the context of the INDECT project, work package 4 is responsible for the Extraction of Information for Crime Prevention by Combining Web Derived Knowledge and Unstructured Data. This document describes the first deliverable of the work package which gives an overview about the main methodology and description of the XML data corpus schema and describes the methodology for collection, cleaning and unified representation of large textual data from various sources: news reports, weblogs, chat, etc.

2 Introduction

This section provides an overview of deliverable 4.1, the list of participants and their roles as well as a thorough description of the annotation schemes used in publicly or under licence available corpora.

The aim of work package 4 (WP4) is the development of key technologies that facilitate the building of an intelligence gathering system by combining and extending the current-state-of-the-art methods in Natural Language Processing (NLP). One of the goals of WP4 is to propose NLP and machine learning methods that learn relationships between people and organizations through websites and social networks. Key requirements for the development of such methods are: (1) the identification of entities, their relationships and the events in which they participate, and (2) the labelling of the entities, relationships and events in a corpus that will be used as a means both for developing the methods.

2.1 Objectives and Results

In this report, we provide an overview and a thorough review of the annotation schemes used to accomplish the above goals. Based on our review, we propose a new annotation scheme able to extend the current schemes. The WP4 annotation scheme is used for the tagging of the XML data corpus that is being developed within workpackage 4. Our general objectives can be summarised as follows:

Review of current annotation schemes for entity resolution and attribute identification
Our first objective is the study and critical review of the annotation schemes employed so far for the development and evaluation of methods for entity resolution, co-reference resolution and entity attributes identification.

Proposal of a new annotation & knowledge representation scheme
Based on the first objective, our second goal is to propose a new annotation scheme that builds upon the strengths of the current-state-of-the-art. Additionally, the new annotation scheme should be extensible and modifiable to the requirements of the project.

2.1.1 Main Objectives

Given an XML data corpus extracted from forums and social networks related to specific threats (e.g. hooliganism, terrorism, vandalism, etc.); an annotation and knowledge representation scheme that should provide the following information:

- The different entity types according to the requirements of the project.
- The grouping of all references to an entity together.
- The relationships between different entities.
- The events in which entities participate.

Additionally the annotation and knowledge representation scheme should be extensible to include new semantic information..

2.1.2 Main Achievements and/or Possible Applications

The main achievements of this work can be summarised as follows:

2.1.2.1 WP4-annotation & knowledge representation scheme

The WP4-annotation & knowledge representation scheme allows the identification of several types of entities, groups the same references into one class, while at the same time allows the identification of relationships and events.

The inclusion of a multi-layered ontology ensures the consistency of the annotation, and allows the satisfaction of the requirements of extensibility and modifiability of the current scheme.

2.1.2.2 WP4-annotation & knowledge representation scheme applications

The WP4-annotation & knowledge representation scheme facilitates the use of inference mechanisms such as transitivity to allow the development of search engines that go beyond simple keyword search. This is accomplished by the use of a multi-layered ontology.

Additionally, the rich annotation offers a benchmark for the evaluation of NLP methods as well as a significant resource for their development and fine-tuning.

2.2 List of participants & roles

This report has been produced by the University of York (UOY), and has been utilized by INNOTEC for the purpose of dissemination (D.9.1)

2.3 Description of Datasets & Annotation Schemes

In this report we focus on the annotation schemes used in a set of 6 publicly or under license available corpora. These datasets/annotation schemes are the following:

- Automatic Content Extraction (ACE)

  The first dataset is the Automatic Content Extraction Dataset (release: LDC2007E63) [2]. This dataset is provided by the Linguistic Data Consortium [1] under license. This dataset has been produced using a variety of sources, such as news, broadcast conversations, etc. Table 1.1 provides an overview of the dataset properties. More importantly, ACE annotation also focuses on co-reference resolution, identifying relations between entities, and the events in which these participate.

- Knowledge Base Population (KBP)

  The annotation scheme in KBP focuses on the identification of entity types of Person (PER), Organization (ORG), and Geo-Political Entity (GPE), Location (LOC), Facility (FAC), Geographical/Social/Political (GPE), Vehicle (VEH) and Weapon (WEA).

  The goal of the 2009 Knowledge Base Population track (KBP) [3] is to augment an existing knowledge representation with information about entities that is discovered from a collection of documents. A snapshot of Wikipedia infoboxes is used as the original knowledge source. The document collection consists of newswire articles on the order of 1 million. The reference knowledge base includes hundreds of thousands of entities based on articles from an October 2008 dump of English Wikipedia. The annotation scheme in KBP focuses on the identification of entity types of Person (PER), Organization (ORG), and Geo-Political Entity (GPE).

- NetFlix

NetFlix [9] is a movie rental site that has started a competition to improve upon its movie recommendation engine. The movie rating data contain over 100 million ratings from 480 thousand randomly-chosen, anonymous Netflix customers over 17 thousand movie titles. It is straightforward that NetFlix focuses on a domain-specific task, hence its annotation is well-suited for this domain.

- WePS-2

The Web People Search (WePS) workshop [4, 5] focuses on two tasks: (1) clustering web pages to solve the ambiguity of search results, and (2) extracting18 kinds of attribute values for target individuals whose names appear on a set of web pages. Similarly to ACE & KBP, WePS annotates entity names, and their attributes, i.e. relationships, birth dates and others.

Source	Training epoch	Approximate size
Broadcast News	3/03 - 6/03	55,000 words
Broadcast Conversations	3/03 - 6/03	40,000 words
Newswire	3/03- 6/03	50,000 words
Weblog	11/04 - 2/05	40,000 words
Usenet	11/04 - 2/05	40,000 words
Conversational Telephone Speech	11/04-12/04	40,000 words

Table 1.1: ACE training corpus statistics for release LDC2007E63

3 Review on current annotation schemes

This section provides a detailed review of the ACE and KBP annotation schemes.

3.1 Automatic Content Extraction Annotation Scheme

The purpose of this section is to present the annotation scheme employed by the Automatic Content Extraction (ACE) project[2]. The vast amount of electronic information, most notably lying around the web, provides a huge resource that can be exploited to enhance the development of natural language understanding applications.

However, in order to take advantage of this potential, it is essential to develop technology that extracts content from human language automatically. This is the objective of the ACE project, i.e. the development of content extraction technology that supports the automatic processing and exploitation of language data in text form. Language data is derived from a variety of sources such as newswire, forums, blogs, etc. The ACE scheme supports a large number of Natural Language Processing (NLP) applications by extracting and representing language content, i.e. the meaning conveyed by the data.

The specific objective of the ACE project is to develop technology to automatically infer from human language data the following:

- The named entities being mentioned in text.
- The relations that exist among the identified entities.
- The events in which the identified entities participate.
- All references to an entity and its properties.

It should also be mentioned that the ACE data sources include audio and imaged data in addition to pure text. In addition to English, ACE has also released datasets for Chinese and Arabic. Based on the above, the ACE project consists of the following four tasks:

1. Entity Detection & Characterization (EDC)

2. Relation Detection & Characterization (RDC)

3. Event Detection & Characterization (EDC)

4. Entity Linking Tracking (LNK)

[2] http://www.itl.nist.gov/iad/mig/tests/ace/ [Accessed 15/06/2009]

In the following sections, we describe each of the tasks in terms of: (1) the language data that they annotate, (2) the categorization scheme employed to organize the annotated data, and (3) the exact annotation used in text. For each task and type of data annotation we provide a number of examples to allow the comprehension of the annotation framework.

3.1.1 Entity Detection & Characterization (EDC)

The goal of the ACE EDC task is the recognition of entities, not just names. This means that all mentions of an entity, i.e. a name, a description, or a pronoun are identified and then classified into equivalence classes. Therefore, co-reference resolution (Entity Linking Tracking task) is important. ACE classifies entities into one of seven main types, which are further divided to more specific subtypes. These five main types are the following:

- Persons (PER)
- Organizations (ORG)
- Location (LOC)
- Facility (FAC)
- Geographical/Social/Political (GPE)
- Vehicle (VEH)
- Weapon (WEA

3.1.1.1 Persons (PER)

This type is used to annotate entities that refer to a distinct person or a set of people. For instance, a person might be specified by his/her name (e.g. George Robertson), occupation (e.g. the lawyer), family relation (e.g. uncle), pronoun (e.g. he), or any combination of these. The **Persons** is further subdivided into the following subtypes:

- **Individual (Tag: PER. Individual)**

  The identified entity refers to a single person. For example[3] and bold its head.

  *[**Silvio Berlusconi**] [The **prime minister** of Italy] Police found[**his** gun].*

- **Group (Tag: Per.Group)**

[3] Square brackets indicate the extend of an Entity Mention

This subtype is used to classify an entity which refers to a group of people unless the group can also be characterized as an Organization or GPE. This not represented by a formal organization (e.g. The ancient Greeks). For example:

[The **Walshes**] family [The **friends** of Arsenal]

- **Indeterminate (Tag: Per. Indeterminate)**

  In cases, where it is impossible to assign an entity to one of the afore mentioned types based on the available context, then the indeterminate type is used.

3.1.1.2 Organizations (ORG)

The mention of an organization or a group of organizations in a given document gives rise to an entity type of Organization. Note that an Organization entity must have been established in a formal manner. Some examples of organizations are firms, government units, sports teams, music groups and others. This entity type is divided into the following subtypes:

- **Government (Tag: ORG.GOV)**

  This subtype refers to entities that are related to governmental affairs, politics, or the state. Note that the entire government of a GPE is excluded from this subtype and should be classified as GPE.ORG as we will see later. This subtype also includes military organizations that are connected to the government of a GPE. Some examples are the following:

  *[The **British navy**] announced yesterday that . . .*

  *[The **ministry of culture**] has funded our research . . .*

- **Commercial (Tag: ORG.COM)**

  This subtype refers to organizations, which primarily focus on providing products or services for profit. Some examples are the following:

  *[**Google's** search engine] is based on PageRank . . .*

  *[**Apple**] announced yesterday the release of its new iPhone . . .*

- **Educational (Tag: ORG.EDU)**

  This subtype refers to organization, which primarily focus on providing educational services. Some examples are the following:

  *[**The University of York**] was founded in . . .*

  *[**The University of Dublin**] has an excellent reputation . . .*

- **Entertainment (Tag: ORG.ENT)**

  This subtype refers to organizations, which primarily focus on providing entertainment services, but excludes giant organizations such as Disney, which is a commercial organization. Some examples are the following:

  *[**The York Theater Company**] decided to increase the number of plays . . .*
  *[**Beatles**] was one of the most famous music groups that . . .*

- **Non-Governmental Organizations (Tag: ORG.NonGov)**

  This subtype includes organizations that are not related to a government or a commercial organization, and whose primary focus is politics (in a broad sense), charity, and advocacy.

It includes several diverse organizations such as paramilitary groups (e.g. [PKK]), political parties (e.g. [Labor Party]), political advocacy groups (e.g. [Palestinian Support Group]), professional regulatory and advocacy groups (e.g. [The Greek Medical Association]), charitable organizations (e.g. [The Red Cross]) and international political bodies (e.g. [The E.U]).

- **Media (Tag: ORG.MED)**
  This subtype includes organization whose primary focus in on the distribution of news or publications. This subtype will include organizations such as BBC, Guardian, National Geographic, etc.

- **Religious (Tag: ORG.REL)**
  This subtype includes organizations that focus on religious issues and affairs. Some examples are the following:

  [The ***Orthodox Church****] was established after the division . . .*
  *[****The Vatican****] aims to have good relations with . . .*

- **Medical-Science (Tag: ORG.SCI)**
  This subtype includes organizations that focus on applying medical care or scientific research. Some examples are the following:

  *[****The London General Hospital****] deals everyday with 1000 of cases . . .*
  *[****NHS****] has designed a program to quit smoking . . .*

- **Sports (Tag: ORG.SPO)**
  This subtype includes organizations that focus on organizing or participating in athletic events. These events can be professional, amateur or scholastic. This subtype also includes groups whose sports are board or card games. Some examples are the following:

  *[****The International Football Federation****] set new rules . . .*
  *[****Manchester United****] has lost the European championship, because . . .*

Finally, it should be noted that many organization might fit to more than one subtype. In these cases, ACE annotators assign organizations to the most specific subtype.

- **Geographical/Social/Political Entities (GPE)**
  This type includes composite entities that consist of a population, a government, a physic al location, and a nation (or province, state, county, city, etc.). All the mentions of these aspects are marked as GPE. For example, in the phrase *the people of U.K*, there are two mentions that are marked, i.e. *people* and *U.K.* This is because these mentions are co referenced, as they refer to different aspects of a single GPE. The government of a country is also treated as a reference to the same entity represented by the name of the country. Thus, the *Greece* and *Greece's government* are mentions of the same entity. Note however that specific units within a government are tagged as organizations. GPE type is divided into the following subtypes:

- **Continent (Tag: GPE.CON)**
  This subtype includes mentions of the entireties of one of the seven continents. Some examples are the following:

  [Many countries in **Africa**] have decided to . . . [**Europe**]

- **Nation (Tag: GPE.NAT)**
  This subtype includes mentions of the entireties of any nation. Some examples are the following:

  Al l people in that flight were [**Polish**]. . . Al l survivors were [**Italian**]. . .

- **State-or-province (TAG: GPE.STA)**
  This subtype includes mentions of the entireties of any state, province, canton of a nation. An example is the following:

  *[**New York's** governor] was elected yesterday . . .*

- **County-Or-District (Tag: GPE.COU)**
  This subtype includes mentions of the entireties of any county, district, prefecture of a nation. Some examples are the following:

  *[**Yorkshire** County] is one of the most popular . . .*

  *[**Kavala** prefecture] is located on the north part of Greece. . .*

- **Population-Center (Tag: GPE.POP)**
  This subtype includes mentions of the entireties of any GPE below the level of GPE.CON. An example is the following:

  *[**York's** mayor] announced yesterday . . .*

- **GPE-Cluster**
  This subtype includes groupings of GPE that can function as political entities (e.g. [the **European Union**]).

It should be noted that: (1) non-political clusters of GPE are marked as Location (e.g. [the **Northern Italy**]), (2) coalitions of governments are marked as Organizations (e.g. [the **NATO**]). Additionally, each GPE entity is associated to a role that can be Person, Organization, Location, or GPE. This judgment depends on the relations that the entity enters into.

3.1.1.3 Locations (LOC)

Places referring to geographical or astronomical regions and do not constitute a political entities give rise to Location entities. For example, the Ouse river, Mountain Everest, or the solar systems are location entities. This type is further divided into the following subtypes:

- **Address (Tag: LOC.ADD)**
  This subtype includes postal addresses or the name of a location (e.g. [**25 WindMill Lane**])

- **Boundary (Tag: LOC.BON)**
  This subtype includes one-dimensional locations such as a border between GPEs e.g. [**The borders**] shared by Greece and Turkey).

- **Celestial (Tag: LOC.CEL)**
  A location which is otherworldly or entire-word-inclusive (e.g.[**The sun**]).

- **Water-Body (Tag: LOC.WAT)**
  Bodies of water (e.g.[The **Ouse river**])

- **Land-Region-natural (Tag: LOC.LAN)**
  Natural locations that are geologically or ecosystemically designated (e.g. [**The Grand Canyon**]).

- **Region-International (Tag: LOC.REGI)**
  Locations that cross national borders (e.g. [The **Eastern Europe**.]

- **Region-General (Tag: LOC.REGG)**
  Locations that do not cross national borders (e.g. [The **eastern Italy**.]

3.1.1.4 Facilities (FAC)

In ACE a facility is defined as a functional, primarily man-made structure. This type includes buildings such as airports, stadiums, factories, museums, prisons, etc. They can be considered as artifacts of the domains of civil engineering or architecture. Facilities are further subdivided into the following:

- **Airport (Tag: FAC.AIR)**
  This subtype includes airports (e.g. [The **Venizelos airport** in Athens . . .].

- **Plant (Tag: FAC.PLA)**
  This subtype includes buildings used for industrial purposes (e.g. [the oil refinery]).

- **Building-or-Grounds (Tag: FAC.BUI)**
  This subtype includes man-made and man-maintained buildings or outdoor facilities (e.g. [the Berlin Wall]).

- **Subarea-Facility (Tag: FAC.SUB)**
  This subtype includes rooms, apartments and other areas that allow a person to live (e.g. [the apartment] I rented. . .).

- **Path (Tag: FAC.PAT)**
  This subtype includes facilities that allow the flow of fluids, energies, persons (e.g. [the telephone lines]).

3.1.1.5 Vehicle (VEH) & Weapon (WEA)

Vehicle (VEH) & Weapon (WEA) are two types which are also included in ACE 2008.

However, the entity task guidelines do not describe these types.

3.1.2 Relation Detection & Characterization (RDC)

The goal of this task is to identify and characterize the relations between two target entities that have been identified in the EDT task. Each relation takes two arguments, i.e. the entities that participate in the relation. Each identified relation has to be assigned one of the seven class types. These types and their subtypes are the following:

3.1.2.1 Physical (Tag: PHYS)

This type includes relations that describe physical proximities of target entities. It is further divided into the following subtypes.

- **Located (Tag: PHYS.Located)**
  This relation captures the location of an entity with respect to another entity. Some examples are the following:

  *PHYS.Located ([The **station** is located at the top of the hill, **the top of the hill**])*
  *PHYS.Located ([The base in London**, London**])*

- **Near (Tag: PHYS.Near)**
  This relation indicates that an entity is near another entity. However, it is neither in that location nor part of it. An example is the following:

  *PHYS.Near ([**The station** is 20 miles north of London, **London**])*

- **Part-whole (Tag: PHYS.Part-Whole)**
  This relation indicates that an entity is part of another one. An example is the following:

  *PHYS.Part-Whole ([A **state** within the US, the **US**])*

3.1.2.2 Personal/Social (Tag: PER-SOC)

This type describes relations between entities of type PER. The order of the arguments does not have any impact on the relations. It is further divided into the following subtypes.

- **Business (Tag: PER-SOC.Business)**
  This relation captures the professional connection that exists between two entities. For example:

  *PER-SOC.Business ([My **lawyer**, **lawyer**])*
  *PER-SOC.Business ([The senator's **secretary**, the **senator**])*

- **Family (Tag: PER-SOC.Family)**
  This relation captures family relations. For example:

  PER-SOC.Family (My lawyer, lawyer])
  PER-SOC.Family (The senator's secretary, the senator])

- **Other (Tag: PER-SOC.Other)**
  All other social relationships that to do not fit into the above subtypes are assigned to PER SOC.Other. For example:

  *PER-SOC.Other ([George's **flatmates**, the **George**])*

3.1.2.3 Employment/Membership/Subsidiary (Tag: EMP-ORG)

This relation includes employment relations between PERs and an ORG or GPE, subsidiary relations between ORGs and GPEs, and membership relations between one of PER, ORG, GPE and an ORG. It has the following subtypes:

- **Employ-exec(s) (Tag: EMP-ORG.Employ-exec)**
  This subtype captures employment relations between persons and organizations with the restriction that the person holds a managerial position such CEO, director, etc. For example:

  *PER-ORG.Employ-exec ([The **CEO** of Google,**Google**])*
  *PER-ORG.Employ-exec ([US **president,US**])*

- **Employ-staff (Tag: EMP-ORG.Employ-staff)**
  This subtype captures employment relations between persons and organizations with the restriction that the person holds a non-managerial position such CEO, director, etc. For example:

  *PER-ORG.Employ-staff ([A web **designer** in Google., **Google**])*

- **Employ-undetermined (Tag: EMP-ORG.Employ-undetermined)**
  In cases, where the context does not provide enough information whether an individual has managerial position or not, the Employ undetermined is used. For example:

  *PER-ORG.Employ-undetermined ([**John** has been working in Google since . . . , **Google**])*

- **Member-of-group (Tag: EMP-ORG.Member-of-group)**
  This subtype includes membership relations. For example:

  PER-ORG.Member-of-group
  *([**John** is a member of the Labors, **Labors**])*

- **Subsidiary (Tag: EMP-ORG.Subsidiary)**
  This subtype characterizes the relationship between a company and its parent company. For example:

  PER-ORG.Subsidiary ([**Google** parent company of YouTube, **YouTube**])

- **Partner (Tag: EMP-ORG.Partner)**
  This subtype characterizes the relationship between a partner companies. For example:

CNN and NBC announced their partnership . . .
*PER-ORG.Partner ([**CNN, NBC**])*

- **Other (Tag: EMP-ORG.Other)**
  Any other collaborative relationship that does not fit into one of the above subtypes is assigned to EMP-ORG.Other

3.1.2.4 Agent-Artifact (Tag: ART)

This type captures the relationship between agentive entities and artifacts. It is divided into the following subtypes:

- **User/Owner (Tag: ART.User-Owner)**
  An agent is the owner or the possessor of an artifact. For example:

  *My house was built five years ago. ART.User-Owner ([**My**, My **house**])*

- **Inventor/Manufacturer (Tag: ART.Inventor-Manufacturer)**
  An agent is the inventor or the manufacturer of an artifact. For example:

  Lary Page, the inventor of PageRank . . .
  *ART. Inventor-Manufacturer ([**Lary Page**, **Larry Page**])*

- **Other (Tag: ART.Other)**
  Any other Agent-Artifact relationship that does not fit into one of the above subtypes is assigned to ART.Other.

3.1.2.5 PER/ORG Affiliation (Tag: Other-AFF)

This type describes the relationship between entities that are not captured by other types. It is further subdivided into the following sub-types:

- **Ethnic (Tag: Other-AFF.Ethnic)**
  This subtype captures the relationship between Person(s) and a group PER to which they belong. For example:

  African-American people . . .
  *Other-AFF.Ethnic ([African-American **people, African-American**])*

- **Ideology (Tag: Other-AFF.Ideology)**
  This subtype captures the relationship between Person(s) and a group PER/ORG to which they belong with the restriction that the group is defined by coherent ideological systems. For example:

  Christian people . . .
  *Other-AFF.Ideology ([Christian **people, Christian**])*

- **Other (Tag: Other-AFF.Other)**
  This subtype should be used in cases where all PER-ORG Affiliation relations do not fit into any of the above categories.

3.1.2.6 GPE Affiliation (Tag: GPE-AFF)

This type describes the relationship between entities of type PER, ORG and GPE, when more than aspect of the GPE is mentioned in the context. It is further subdivided into the following subtypes:

- **Citizen/Resident (Tag: GPE-AFF.Citizen)**
  This subtype describes the citizen or resident relationship between a PER and a GPE. For example:

  US athlete Michael Jordan . . .
  *GPE-AFF.Citizen ([US **athlete, US**])*

- **Based-in (Tag: GPE-AFF.Based-In)**
  Given that organizations are not always located in the GPE in which they are based, ACE distinguishes between the physical locations of an ORG with their GPW of origin. For example:

  Google Zurich focuses on the development . . .
  *GPE-AFF.Based-In ([**Google** Zurich, **Zurich**])*

- **Other (Tag: GPE-AFF.Other)**
  This subtype is used for GPE affiliations that do not fit to any of the above subtypes.

3.1.2.7 Discourse (Tag: DISC)

A Discourse relation captures part-whole or membership relations, which are established only for the purposes of the discourse. The group entity referred to is not an official entity relevant to world knowledge. For example:

Many of these people . . .
*DISC ([**Many** of these people, **people**])*
*DISC ([**Each** of whom, **whom**)]*

3.1.3 Event Detection & Characterization (EDC)

The goal of this task is to identify and characterize events according to five predefined types. Each event is tagged by its textual anchor, full extend, and participating entities. For each event type there is a salient entity. A salient entity can be the object of the event (Object Salient Events), or the agent of the event (Agent Salient Events). Table 2.1 shows this classification. In the following examples, square brackets are used to denote the extend of an event, curly brackets are used to denote the anchor of an event, while parenthesis are used to identify the salient entity.

Event Type	Salient Entity Role
MOV	Object
BRK	Object
MAK	Object
GIV	Object
INT	Agent

Table 2.1: Event type & salient entity roles Many of these people . . .

Object Salient Events
As it has been mentioned in Object Salient Events, the EDT entity filling the object role is the focus of the event. There are four types of Object Salient Events:

- **Destruction/Damage (BRK)**
  An event is classified as BRK, when the salient entity is damaged, destroyed or killed. For example:

  [Last year, (5 people) were {killed} by the terrorist attack in Gaza.]

- **Creation/Improvement (MAK)**
  An event is classified as MAR, when the salient entity is created, improved or born. For example:

[(Google) was {founded} by Larry Page and Sergey Brin.]

- **Movement (MOV)**
  An event is classified as MOV, when the salient entity is moved. For example:

  [(Google) {moved} to Mountain View.]

- **Transfer of Possession or Control (GIV)**
  An event is classified as GIC, when the salient entity changes with respect to possession or control. For example:

  [(He) was {arrested} with the charge of possessing weapons.]

Agent Salient Events
As it has been mentioned in Agent Salient Events, the EDT entity filling the agent role is the focus of the event. There is only one subtype for this type which is the following:

- **Interaction of Agents (INT)**
  An event is classified as INT, when the salient entities are agents engaged in some kind of interaction. Note that an entity can be an agent if its type is PER, ORG, or GPE. For example:

  [(Five thousand people) {demonstrated} in Athens, protesting against the death of a teenager . . . .]

3.1.4 Entity Linking Tracking (LNK)

The goal of the Entity Linking task is to group all references to an entity and its properties together. While an Entity is an object or set of objects in the world that can be referenced by their name, a nominal phrase, or a pronoun, a Composite Entity results from linking an Entity to all attributive mentions of its properties.

Entity Grouping

All specific and generic entities are linked with the predicates and other attributive mentions that ascribe properties to them. This ensures that each composite entity consists of a set of strings, which either refer to or describe a given entity in text. The following relations are examined for entity linking.

- **Predicate complements**
  In cases where a property is ascribed to an entity via an asserted predicate complement, the attributive mention is linked with the entity it describes. For example:

  [London] is [a very popular destination]

- **Apposition**
  In cases of apposition, the first element is a specific reference to the entity, while the second element is an attributive mention. For example:

  [London], [a very popular destination]

- **Premodifiers**
  All premodifiers are attributive; hence they are linked with referential entities when they ascribe a property that is derived from an entity. For example:

  [Greece] is a very popular destination, and [Greek] islands are famous for their clean sea.

The specific referential entity Greece and the attributive mention Greek will be linked, since Greek is ascribing Greece attributes to islands.

Cross-type Metonymy
Cross-type Metonymy can happen when a composite entity consists of EDT entities that can be assigned to different EDT types depending on the context. One example is that of ORGs and the FACs they occupy. While in the EDT stage these two characteristics are tagged separately (ORG & FAC) depending on context, in this stage group entities of different types are grouped together into a composite entity by creating links between them when they refer to different aspects of the same underlying object. For example:

[The White House] announced yesterday that . . .

[John Smith reports from the White House park] . . .

In this example, the first mention of White House is of type EDT.ORG. However, the second mention is of type EDT.FAC. Each of these mentions will be linked, since they evoke different aspects of the same underlying entity.

3.2 The KBP annotation scheme

The goal of the KBP track at the 2009 Text Analysis Conference is to evaluate the ability of automated systems to discover information about named entities and to incorporate this information in a knowledge source [3]. KBP consists of be two related tasks: Entity Linking, where names must be aligned to entities in a knowledge base, and Slot Filling, which involves

mining attributes of entities from text.
In contrast to ACE, KBP focuses on the following types of entities:

- Person (PER)
- Organization (ORG)
- Geo-Political Entity (GPE)

The description of the KBP scheme does not provide any details regarding the categorization of the top-level types to more specific ones. However, as in the ACE evaluation, GPEs include inhabited locations with a government such as cities and countries. Wikipedia infoboxes are the basis for the reference knowledge base; An infobox is a data structure that allows the description of a target entity through a set of desired attributes called slots. There is one generic infobox for each entity type.

Table 2.2 shows these generic infoboxes and their slots that include the attributes of entities. As it can be observed, KBP provides a richer scheme in terms of entities attributes and relations than ACE.

On the other hand, ACE provides a clear classification of relation types, which ensures consistency and avoids duplications. In the next section, we present the advantages and disadvantages of using infoboxes as a knowledge representation scheme as opposed to having a fixed set of relations or events. Based on that discussion, we propose an extended version of ACE that includes infoboxes in the next chapter.

Organization	Geo-Political	Entity
per:alternate_names	org:alternate_names	gpe:alternate_names
per:date_of_birth	org:political/religious_affiliation	gpe:capital
per:age	org:top_members/employees	gpe:subsidiary_orgs
per:place_of_birth	org:number_of_employees/members	gpe:top_employees
per:origin	org:members	gpe:political_parties
per:date_of_death	org:member_of	gpe:established
per:place_of_death	org:subsidiaries	gpe:population
per:cause_of_death	org:parents	gpe:currency
per:residences	org:founded_by	
per:schools_attended	org:founded	
per:title	org:dissolved	
per:member_of	org:headquarters	
per:employee_of	org:shareholders	
per:religion	org:website	
per:spouse		
per:children		
per:parents		
per:siblings		
per:other_family		
per:charges		

Table 2.2: Slot names for the three generic entity types

3.3 Other annotation schemes

As it has already been mentioned in the first chapter, NetFlix is a movie rental site that has started a competition to improve upon its movie recommendation engine. The movie rating data contain over 100 million ratings from 480 thousand randomly-chosen, anonymous NetFlix customers over 17 thousand movie titles. The ratings are on a scale from 1 to 5 (integral) stars. Training data consist of a file for each movie. The first line of each file contains the movie id followed by a colon. Each subsequent line in the file corresponds to a rating from a customer and its date in the following format: CustomerID, Rating, Date.

In the introductory section we also mentioned that the WePS workshop [4, 5] focused on two tasks, i.e. clustering web pages to solve the ambiguity of search results, and extracting 18 kinds of attribute values for target individuals whose names appear on a set of web pages.

The WePS development data consist of 47 ambiguous names and up to 100 manually clustered search result for each name. The test data consists of 30 dataset where each one corresponds to one ambiguous name. The sources used to obtain the names were Wikipedia biographies, ACL'08 committee members and US census data. In average, there are 18.64 different people per name, but the predominant person for a given name owns half of the documents. A sample cluster set for target Abby Watkins is given below:

```
<?xml version="1.0" encoding="UTF-8"?>
    <clustering name="Abby Watkins">
        <entity id="7">
            <doc rank="111" />
        </entity>
        <entity id="2">
            <doc rank="81" />
        </entity>
        <entity id="0">
            <doc rank="21" />
        </entity>
        <entity id="14">
            <doc rank="99" />
            <doc rank="36" />
            <doc rank="52" />
        </entity>
     </clustering>
```

For the second task the organizer distributed the target Web pages in their original format, (i.e., html), and the participants were to expected to extract attribute values from each page. The individual names associated with a particular page were given, and the attribute values for that person should be extracted. Web pages containing multiple individuals sharing the same name will not be given. Table 2.3 lists the attributes used in the task and annotation scheme

ID	Attribute Class	Examples of Attribute Value
1	Date of birth	4 February 1888
2	Birth place	Brookline, Massachusetts
3	Other name	JFK
4	Occupation	Politician
5	Affiliation	University of California, Los Angeles
6	Work	The Secrets of Droon
7	Award	Pulitzer Prize
8	School	Stanford University
9	Major	Mathematics
10	Degree	Ph.D.
11	Mentor	Tony Visconti
12	Location	London
13	Nationality	American
14	Relatives	Jacqueline Bouvier
15	Phone	+1 (111) 111-1111
16	FAX	(111) 111-1111
17	Email	xxx@yyy.com
18	Web site	http://nlp.cs.nyu.edu

Table 2.3: Definition of 18 attributes of Person at WePS-2

3.4 Discussion

It is apparent that the annotation scheme of ACE provides a rich scheme for the identification, grouping of entities and the discovery of the relations and events, in which they participate. However, ACE does not include a knowledge base, which would enhance its extensibility and modifiability according to the domain or genre of interest.

The extensibility of ACE to specific domains of interest is essential, since it would allow the development of methods focusing on domain-specific threats, such as hooliganism, vandalism, terrorism, and other.

KBP has a significant advantage over ACE's annotation and knowledge representation scheme in that it can be easily extended. This is a consequence of the use of Wikipedia infoboxes, in which one can introduce new slot names describing attributes, relations, or events related to an entity.

However, infoboxes are not the ideal representation scheme, since they can introduce duplication and loss of integrity. This is verging on something that should be classified as a major problem with this representation. The following example illustrates this problem:

Infobox: Bill Clinton	**Infobox: Barack Obama**
Name: Bill Clinton	Name: Barrack Obama
Date of birth:19/08/1946	Date of birth:04/08/1961
Office: President of the United States	Office: President of the United States
Spouse: Hillary Rodham Clinton	Spouse: Michelle Obama
Education: BSc in..., PhD in....	***University Degree: BSc in..., MSc in...***
Website: http://www....	***URL: http://www....***
Children: Chelsea Clinton	Children: Malia Ann Obama , Natasha Obama

Table 2.4: Two example of infoboxes

In Table 2.4, we observe that although both of the entities refer to presidents of the United States, the corresponding infoboxes differ in two slots, i.e. *education-University Degree*, and *website-URL*. This is due to the each slot pair refers to the same underlying concept. For example *education-university degree* refers to the education someone has received, while website-*URL* refer to his/her official website. This inconsistency has been caused by the same property that offers extensibility, i.e. the ability to add new slot names in the created infoboxes.

3.5 Summary

To summarize, this section has provided an overview of the annotation & knowledge representation schemes used in ACE, KBP, NetFlix and WePS. It is apparent that ACE provides a rich scheme, which however is not easily extensible or modifiable, as it lacks structural relationships between objects of interest, while at the same time a knowledge representation scheme is absent.

In contrast, KBP is essentially a subset of ACE in terms of annotation. However, KBP uses a knowledge base and Wikipedia infoboxes as a means to represent knowledge. This allows having an easily extensible and modifiable scheme, yet it introduces duplications and does not ensure integrity. In the next section, we aim to overcome the above limitations by proposing a extended annotation scheme of ACE, which includes the use of an ontology.

4 Design of a new annotation scheme

In this chapter we outline the deficiencies of KBP and ACE, proposing an extension of ACE annotation as the WP4 annotation and knowledge representation scheme. We also argue that a clear and consistent ontology design is a necessity for any application that requires sophisticated search and reasoning and for overall efficiency in knowledge management. We also propose the use of the Proton ontology [8]. The choice of this ontology was motivated by the fact that this ontology already conforms to the ACE annotation guidelines.

4.1 Methodolgy on data collection

D4.1 aims to focus on analysis of security related data from websites, blogs, chats and other social medium. The project aims to analyse data related to hooliganism, terrorism and other types of crime. The AGH (Prof. Wieslaw Lubaszewski's) team has initiated the task of data collection. This section describes the ongoing effort and the methodology employed. It does not include the actual data as this is currently being collected. The current effort is directed towards collecting data on football hooliganism and sale of human organs. In parallel to this, the Ostrava team (Mr Adam Nemcek) has also started work on data collection on similar topics.

The current data collection activity follows the following methodology:

- Only highly relevant data will be collected to ensure that machine learning systems trained using the data will not be swamped by noise.
- The data will be multi-lingual covering a number of different languages. Currently, data is being collected in Polish and Czech.
- Specialised crawlers will be used to help with 1. and 2. and to lessen the need for manual filtering. Both Ostrava and AGH already have built their specialised crawlers. In addition, open source crawlers are also available.
- The subset of the collected data will be annotated using the annotation scheme described in this report. This annotation will be detailed in that it will identify all relevant potential threats, the participants, the locations, the time and connections between entities involved. End users (i.e. the police) will be used to verify the correctness of the annotation where it is necessary.

4.2 Data cleaning methodology

Data from websites, blogs and social networks especially user forums etc. do not always follow strict HTML standards. These are usually ill formed and usually requires cleaning and preprocessing before it becomes usable by any natural language processing pipeline. However, manual cleaning of such data is neither feasible or acceptable as NLP systems developed within the project need to be robust enough to handle such data.

For the above reasons, we propose to employ standard supervised machine learning methods to automatically convert ill-formed data into a well formed corpus.

For example, given an ill-formed HTML such as the following:

```
<html>
	<title>Blog example</title>
	<body>
		<p>
			<strong> johnBam  </strong>
			<strong>12 July 2009, 11.59 GMT </strong>
		</p>
			Italian officials say two train cars filled with liquefied natural gas have
			derailed and exploded in western Italy, killing at least 14 people.

		</p>
	</body>
</html>
```

a human annotator will manually convert the above into its correct form:

```
<html>
	<title>Blog example</title>
	<body>
		<p>
			<strong>Sender:  johnBam  </strong>
			<strong>Date: 12 July 2009</strong>
			<strong>Time: 11.59 GMT </strong>
		</p>
		<p><strong>Text:</strong>
			Italian officials say two train cars filled with liquefied natural gas have
			derailed and exploded in western Italy, killing at least 14 people.

		</p>
	</body>
</html>
```

It can be observed that the tags *Sender, Date, Time, Text* have been inserted in the HTML code to allow the recognition of entities, dates, and text within a blog entry.

The above pair constitutes a single training example. A number of such pairs will be collected to form a training set for a supervised machine learning system such as an SVM. The task of the SVM is to predict the location of different tags (e.g. sender, recipient, posting date etc.) at specific points in the input. This can be formulated as a binary decision problem. A separate

SVM will need to be trained for each tag. And, SVMs can be used in a pipeline to generate all the missing tags.

4.3 WP4 annotation & knowledge representation scheme

The aim of the new annotation scheme is build upon the strengths of ACE annotation scheme and the KBP annotation & knowledge representation scheme. As mentioned in section 2.2, ACE provides a clear classification of relation types, which ensures consistency and avoids duplications.

This should be the primary characteristic of the new annotation scheme. Secondly, ACE annotation already defines a subclass relationship. Wikipedia infoboxes which are used as knowledge bases for KBP are a set of subject-attribute-value triples that lists the key aspect of the articles subject. However using infoboxes as a knowledge representation scheme has the following disadvantages:

- Multiple templates exists for the same class
- Multiple attribute names for the same property
- Attributes lack domains or datatypes

However the infobox classes and attributes can be mapped to corresponding ACE entity and relation annotation scheme. So we can view KBP as a subset of ACE. However for the purpose of this project combining the good features of both annotation schemes seems to be the way ahead.

ACE has clearly defined guidelines for events which the KBP annotation does not address. Meanwhile the infoboxes can be easily extended. However there is no clear ontology defined by these schemes. So an ontology based upon ACE annotation scheme should be implemented.

The need of a better defined ontology is necessary for the following reasons:

Query capabilities
One key advantage of using an ontology is that we can go beyond keyword queries and ask SQL like queries. Ontologies allow various kinds on inference mechanisms such as transitivity to allow sophisticated queries. For example to answer **Which person got the golden boot at 2006 FIFA world cup?** we might have to realize that **footballer** is a subclass of person.

Extensibility
An ontology can be a catalyst for acquisition of further knowledge, largely automated maintenance and growth of the knowledge base. As the knowledge is changing and ever growing, automated extension is a desired characteristic which could be built on to the ontology. There are lot of existing ontologies that use automated knowledge acquisition or extension based on the current ontology relations. For example Gene Ontology (GO) [6] generates more detailed concepts from existing GO concepts by utilizing syntactic relations among the existing concepts.

For example, the hyponymy relation between two concepts *chemokine binding* and *C-C chemokine binding* can be inferred from the hyponymy relation between the subconcepts *chemokine* and *CC chemokine*. In other words, one way to expand an ontology is build upon the relationship between the terms in the existing ontology based on syntactic, dependency and semantic information extracted from the original text containing these terms.

Another such example is the CROSSMARC [7] ontology in which new instances for the existing concepts are learned from domain specific corpus using machine learning approaches. Initially the domain specific corpus is annotated with existing concept instances automatically using the existing ontology. To identify new instances a single Hidden Markov Model (HMM) is trained for each set of instances of a particular concept. HMM parameters are calculated from the annotated domain specific corpus using maximum likelihood estimation. Simply put the HMM learns the context in which the instances occur and use it to detect new instances belonging to the training instance concept.

Expressivity

It can be an enabler for semantic search on the web, for detecting entities and relations in web pages and reasoning about them in expressive logics. For example probabilities can be attached to the concepts and properties during ontology building thus allowing us to reason using probabilistic logic. Such kind of extension reduces the problem in reasoning when only partial information about the concept or the instance exist in the ontology and allows reasoning with partial and imprecise information.

4.3.1 Ontology structure

The central idea is to create an ontology compatible with ACE annotation. This can be done, if the top layer of the ontology reflects the entities defined in ACE. In addition to the ACE entities the top layer also contains a separate class for events and other properties we are interested in such as TimeInterval to denote some timestamp.

An ontology that satisfies the above is Proton ontology. It was developed to be complaint with ACE annotation scheme among others. Proton[4] is divided into four modules: system, top, upper and knowledge management. Figure 3.1 shows the four modules with the classes it contains.

- **System module**
  It introduces the key class entity which can have aliases. Basically this module is used to maintain knowledge that needs to be hard coded for ontology based applications.

- **Top module**
  This module contains the most general class descriptions, about 20 classes. Most of the classes chosen are domain independent with an aim to be able to link existing ontologies.

- **Upper module**
  This module contains more general classes of entities for example various sorts of organizations and a comprehensive range of locations.

[4] http://proton.semanticweb.org/

- **Knowledge management module**
  It contains 38 classes of slightly specialized entities that are specific for typical knowledge management tasks and applications.

As can be observed in Figure 3.1, all of the ACE entity types are incorporated in the top module. Proton was designed to be general purpose and domain independent. The top layer starts with very basic entity classes:

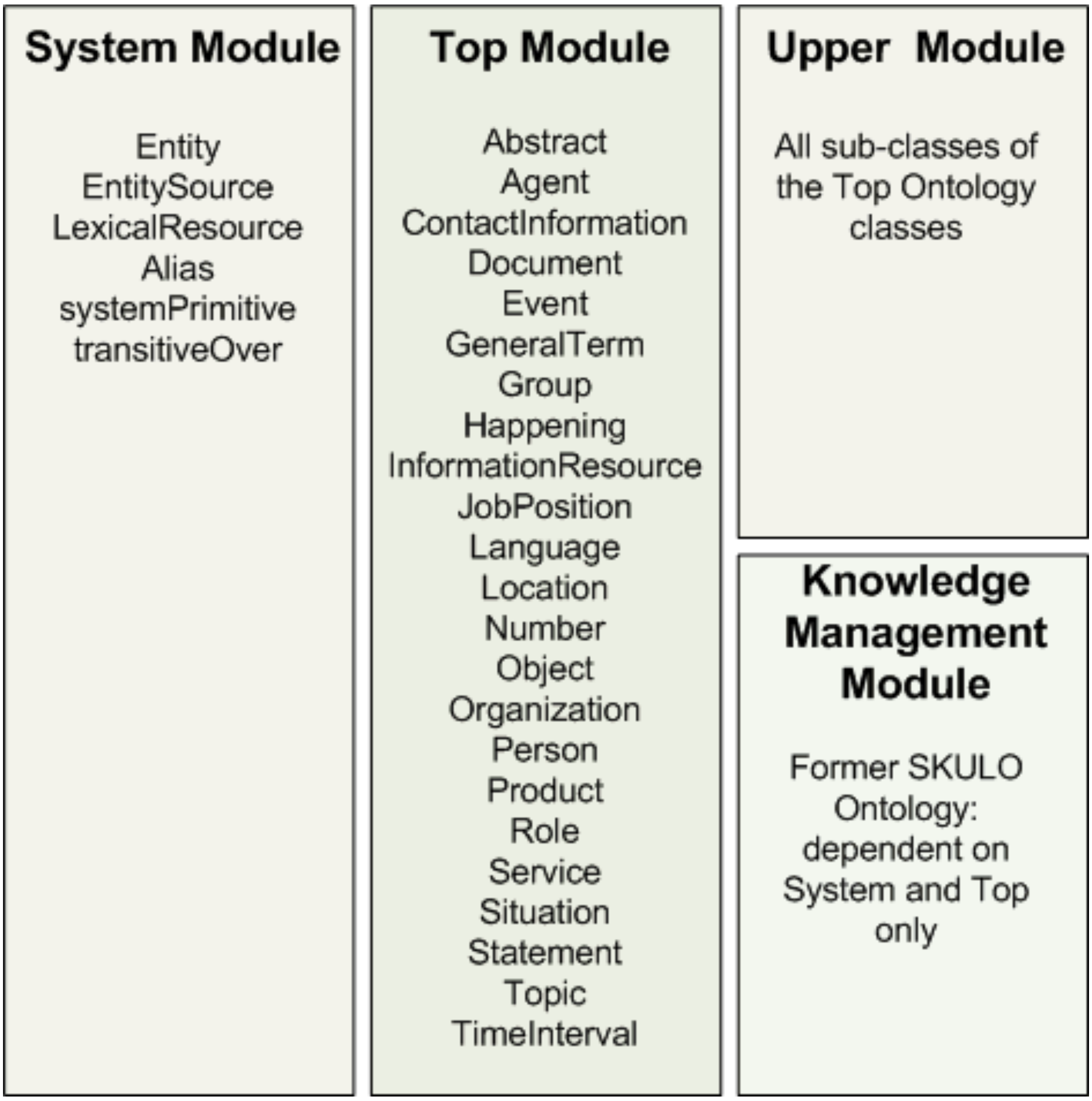

Figure 3.1: Proton Modules

- Object - agents, locations, vehicles etc.
- Happening - events and situations
- Abstract - Abstractions that are neither object or happening

These are further specialized into generally defined entities: meetings, military conflicts, employment (job) positions, commercial, government, and other organizations, people, and various locations. It also covers numbers, time, money, and other specific values.

Additionally, the featured entity types have their characteristic attributes and relations defined for them (e.g. subRegionOf property for Locations, has Position for Person-s, locatedIn for Organization-s, hasMember for Group-s, etc.). Specialization of the classes is achieved with the help of upper layer. For example mountain as a specific type of location and user as a subclass of agent. Separating the ontology into two layers allows for domain specific extensions.

The top module contains the most general classes as per the requirement of the project. The subclasses of these classes belong to the upper module. For example the top class happening includes the subclasses event and situation. Situation is further specialized into jobposition and role. Figure 3.2 shows the top module classes.

The design is an object oriented design. The subclasses inherit the properties from its super classes. For example *person* inherits properties from *agent* and *object*. Apart from the inherited properties, it also has its own properties such as hasPosition (this relates entity person to jobPosition) and hasRelative (this relates person to another one). In this case the hasRelative relationship is of bidirectional many-to-many type. Figure 3.3 shows the specialization of this relationship.

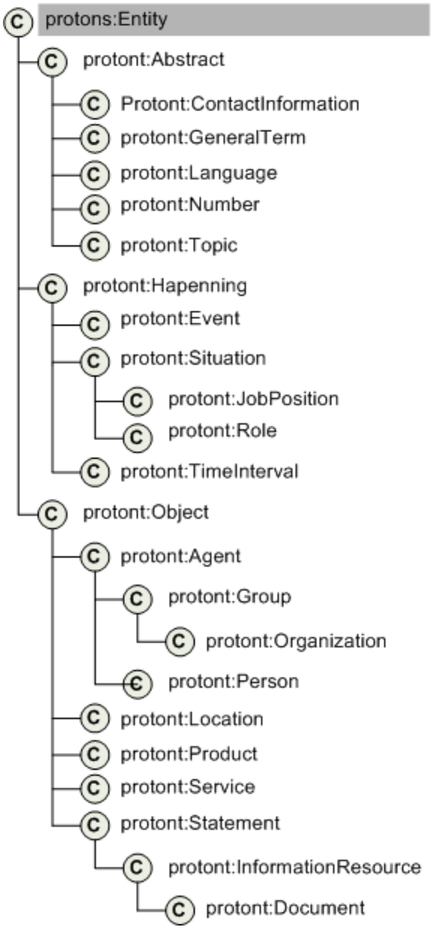

Figure 3.2: Top module classes

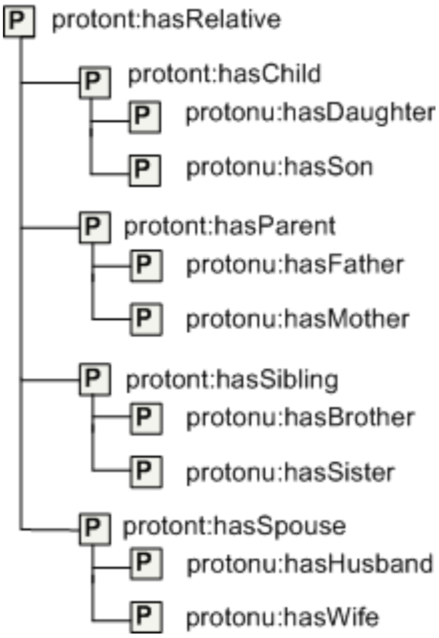

Figure 3.3: Hierarchy of hierarchy of relations

Proton architecture also incorporates events into the top layer. The top layer class *happening* has *event*, *situation*, *timeinterval* and *jobposition* as its subclasses. This class thus incorporates both static (situation) and dynamic happenings (event).

Dynamic events include subclasses like *accident*, *military conflict* and *sports event*. Static events include holding a position like *board member* or *manager*. The rationale is that both static and dynamic event has a temporal marker associated with it, for example a sport has start time and end time. Building an ontology in such a way allows user to search, for example, *all U.K. prime minister before 1980*. The knowledge management module further contains specialization that are specific to knowledge management task. For example the top class *agent* contains *informationsource* subclass that belongs to the knowledge management module. The instance *e-commerce* of *informationsource* contains collection of documents relating to activities and entities concerning electronic commerce.

4.4 Example of annotation & ontology extension

This section provides an example of the WP4 annotation and knowledge representation scheme on three texts extracted from the web. The first text fragment is a weblog on hooliganism; the second is a news report on violent events between hooligans of UK football teams, while the third is a partial transcript of a conversation between terrorists.

The goal in this section is to demonstrate the feasibility of extending the ACE annotation scheme and the associated ontology to new genres. In the following example, entities and their corresponding references are annotated with their corresponding ACE tag.

4.4.1 Weblog on hooliganism[5]

In this weblog no domain-related events have been identified. The tagged entities are shown below.

hejsansvejsan87 [PER.Individual] (6/10/2008, 13:32)
Barca 4 life!
Manchester [ORG.SPO] you suck **cameldick** [ORG.SPO]. **La liga** [ORG.SPO]=THE BEST LEAGUE IN THE WORLD!!!

reilly979 [PER.Individual] (6/10/2008,13:40)
if it sucked it wouldn't be rated 5 stars

hpsjalgpallur [PER.Individual] (6/10/2008,13:45)
this vid sucks because its mostly about **ronaldo** [PER.Individual] and **messi** [PER.Individual] ....

thriller12312 [PER.Individual] (6/10/2008,13:50)
messi [PER.Individual]

soccerismylife1994 [PER.Individual] (6/10/2008,13:53)
messi [PER.Individual] is the best!!
ronaldo [PER.Individual] sucks!!

elegyrulz [PER.Individual] (6/10/2008,13:56)
Lionel [PER.Individual] is the best!

brothering1 [PER.Individual] (6/10/2008,14:00)
I am wondering when will **Barcelona's** [ORG.SPO] arses will get kicked by **Real Madrid** [ORG.SPO] now...

amitpetra [PER.Individual] (6/10/2008,14:10)
Barca [ORG.SPO] forever
Messi [PER.Individual] >>>>>**gaynaldo** [PER.Individual]

foroeste [PER.Individual] (6/10/2008,14:30)
FC Barcelona [ORG.SPO] >>>>>>**manchester** [ORG.SPO]
but **C.ronaldo** [PER.Individual] >>>>>**messi** [PER.Individual]

[5] http:// www.youtube.com/watch?v=dv4DkGK5zNA

4.4.2 News report[6]

Among the most serious incidents reported to the (**National Criminal Intelligence Service** [ORG.GOV]) **NCIS** [ORG.GOV] were:

July 2008: **Glasgow Rangers** [ORG.SPO] v **Shelbourne** [ORG.SPO]. **Police baton** [ORG.GOV] charged 150 **Rangers supporters** [PER.Group] who were trying to attack fans of the **Irish club** [PER.Group].

August 2008: **Norwich City** [ORG.SPO] v QPR [ORG.SPO]: Twenty **supporters from both sides** [ORG.SPO] involved in bottle throwing in a **Norfolk** [LOC.ADD] pub. One person arrested.

30 September 2008: **Norwich City** [ORG.SPO] v **Birmingham City** [ORG.SPO]: Twenty **Birmingham fans** [PER.Group] sprayed rival supporters with **CS gas** [WEA] and attacked them with bar stools in a pub.

- **Identified Events**
  - E1: [Police baton {*charged}* 150 **Rangers supporters** who were trying to attack fans of the Irish club.]
  - E2: [150 Rangers supporters who were {*trying to attack*} **fans of the Irish club**.]
  - E3: [150 **Rangers supporters** who were {*trying to attack*} fans of the Irish club.]
  - E4: [Twenty **supporters from both sides** involved in {*bottle throwing*} in a orfolk pub.]
  - E5: [Twenty **Birmingham fans** {*sprayed}* rival supporters with CS gas]
  - E6: [Twenty Birmingham fans {*sprayed}* **rival supporters** with CS gas]
  - E7**: [**Twenty **Birmingham fans {**attacked} them with bar stools in a pub**]**

Event ID	Event Type
E1	*Hooliganism.Severe*
E2	*Hooliganism.Severe*
E3	*Hooliganism.Severe*
E4	*Hooliganism.Critical*
E5	*Hooliganism.Critical*
E6	*Hooliganism.Critical*
E7	*Hooliganism.Critical*

[6] http://news.bbc.co.uk/1/hi/uk/222225.stm

Table 4.1 Event types for identified events

A new sub-class of *event* say *hooliganism* can be introduced to handle events related to hooligan activities. Hooliganism can be further specialized into events indicating the seriousness of the event, for example: minor, severe, critical and others. Let us assume that the events E1 to E3 are less harmful than E4 to E7, since the agents in E1 to E3 did not actually execute their attack. Their corresponding types are shown in Table 4.1.

The extension of the ontology is straightforward, since the *PROTON* already defines an event class. *Hooliganism* and its subclasses (children) can be added under the *event* node in *Proton*. This means that the top layer remains the same, while the new subclasses can be directly added to the *upper layer* in the *PROTON* hierarchy.

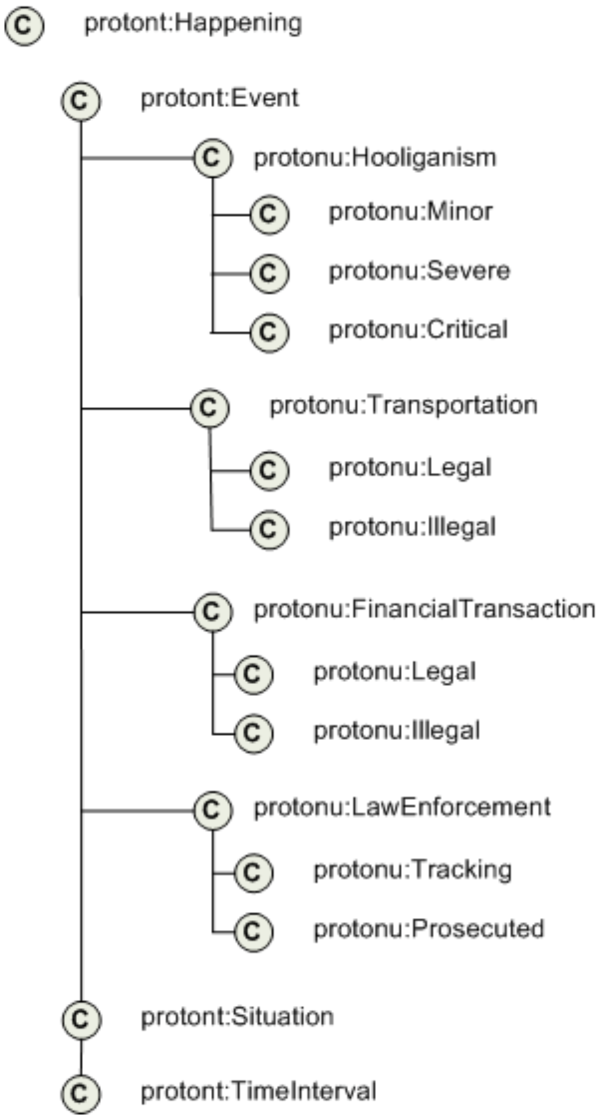

Figure 3.4 Proton Extend Event Hierarchy

Additionally, each of the identified event types in the ontology can be assigned a variety of attributes, which indicate for example whether an event was completed or not, or weights illustrating the severity of the event. The latter allows the development of methods which reason using probabilistic logic.

4.4.3 Terrorist chat[7]

Shazad Tanweer [PER.Individual]: Any extra risks getting into **Pakistan** [GPE.NAT] ?

Omar Khyam [PER.Individual]: We had five **Bengalis** [GPE.NAT] last year. Guess how **we [PER.Group]** got **them** [GPE.NAT] in. From **Bangladesh** [GPE.NAT] all the way across **India** [GPE.NAT] into **Pakistan**[GPE.NAT]... **we** [PER.Group] bribed the guy [PER.Individual]. You know when you [PER.Individual] go to the check-in, it would all be set up.

Mohammed Siddique Khan [PER.Individual]: Going through the airport - normal tickets.

Omar Khyam[PER.Individual]: Yeah, just walk straight through bruv normal, just act as if you are a **Pakistani** [GPE.NAT].

Shazad Tanweer [PER.Individual]: I live in **Faisalbad** [GPE.NAT]

Omar Khyam [PER.Individual]: That's not a problem

Omar Khyam [PER.Individual]: All right **bruv** [PER.Individual]. Get your parents to pick you up. Or your family ... And that way you will breeze through the airport seriously. Even if **they** [ORG.GOV] are following **you** [PER.Individual] - it doesn't really count. Chill out, proper chill out ... until **we** [PER.Group] contact you and then we'll pick **you** [PER.Individual] up.

- **Identified Events**
  - E1: [Guess how **we** {got} them {in}. From Bangladesh all the way across India into Pakistan]
  - E2: [Guess how we {got} **them** {in} From Bangladesh all the way across India into Pakistan]
  - E3: **We** {bribed} the guy.
  - E4: We {bribed} the **guy**.
  - E5: [when **you** {go to the check-in}, it would all be set up.]
  - E6: **[**Even if **they** {are following} you]
  - E7: **[**Even if they {are following} **you**]

Event ID	Event Type
E1	Transportation.Illegal
E2	Transportation.Illegal
E3	FinancialTransaction.Illegal

[7] http://www.channel4.com/news/articles/society/law_order/mi5+transcript+of+bombers+conversation/491157

E4	FinancialTransaction.Illegal
E5	Transportation.Legal
E6	LawEnforcement.Tracking
E7	LawEnforcement.Tracking

Table 4.2 Event types for identified events

In the same vein as in the previous example, we can extend our ontology with different types and subtypes of events. For example, it is apparent that the first two events refer to an illegal transportation, the next two refer to illegal financial transactions, the fifth refers to a legal transportation, and finally the last two refer to law enforcement activities. The corresponding event types which extend the *PROTON* ontology are shown in Table 4.2

4.5 Mapping of publicly available datasets to WP4 annotation scheme

Since the ontology we propose will be based around ACE annotation scheme, we can easily incorporate publically available datasets whose annotations either are directly compliant to ACE or can be mapped to one. In cases where the new dataset has new entity, it can be plugged into the most relevant position in the hierarchy in the ontology. Suppose we did not have a specific category for the entity "asteroids". Since the ontology design starts with very general (system module) and spans to specific instances (upper module) we could still plug "asteroids" as an instance of "object" class. At worst case we can fit any new entity to the "entity" class in the system module.

- NetFlix mapping

  Regarding NetFlix, we can view the data as stating **X commented on movie Y**. In case of the ontology movie would fall under the class **movie** subclass of **mediaproduct** which itself is a subclass of **product**. **Comment** meanwhile is a static event since it was given at a specific time by **X**.

- WePS mapping

  Similarly with the WePS dataset we can view the annotation as stating **X owns document Y**. Since documents are clustered for each name, we can visualize that person owning that web page (artifact).

- KBP mapping

  Finally, as we have already mentioned, KBP can be considered as a subset of ACE, hence the infobox name slots can be easily mapped to their corresponding events or relations included in *PROTON.*

4.6 Summary

It is well understood that KBP annotation has problems regarding consistency and clarify in its definitions. ACE annotation has the basic characteristic that we look for a clean and consistent design. However the knowledge we are trying to maintain can change and evolve over time. As a result an extensible framework for knowledge representation is essential. A multi-layered ontology such as Proton seems to be the way forward. An ontology additionally allows us the use of more powerful and expressive queries.

Multi layered architecture allows the flexibility for the ontology to cater to specific application needs. Furthermore Proton ontology already incorporates ACE annotation scheme. Mapping ACE and KBP annotation scheme onto an ontology is achieved by carefully selecting the top layer classes. Mapping NetFlix and WePS dataset are quite trivial as mentioned earlier.

5 Conclusions

This report has provided a thorough overview of the current-state-of-the-art on the annotation schemes employed for the identification of entities and the attributes that characterize them. The survey part focused on the annotation schemes used publicly and under license available datasets.

In particular, we presented the ACE scheme, which annotates a number of different entity types, relations between them and the events, in which they participate. Following that we presented the KBP annotation and knowledge representation scheme, which in terms of annotation coverage can be considered as a subset of ACE. Additionally, two smaller annotation schemes were discussed, i.e NetFlix and WePS-2.

Based on the critical survey we proposed a new annotation & knowledge representation scheme that extends ACE, so that the new annotation scheme has the following properties:

- It is extensible, in order to fit to the requirements of the project.
  This is particular useful in the early stages of the project where the requirements are not fully specified.
  Extensibility is achieved by using an ontology, which allows the addition of new entities, relations, and events, while at the same time avoids duplication and ensures integrity (as opposed to the KBP scheme).

- It allows a search engine to go beyond simple keyword queries by exploiting the semantic information and relations within the ontology.

- It allows the use of expressive logics and becomes an enabler for detecting entity relations on the web.

6 Bibliography

[1] Linguistic data consortium. http://www.ldc.upenn.edu/ - [Accessed:17/06/2009], 1992.

[2] Automatic content extraction 2008 evaluation plan (ace08), assessment of detection and recognition of entities and relations within and across documents.
http://www.nist.gov/speech/tests/ace/2008/doc/ace08- evalplan.v1.2d.pdf - [Accessed:17/06/2009], April 2008.

[3] Task description of knowledge base population track. http://apl.jhu.edu/paulmac/kbp/090601-KBPTaskGuidelines.pdf - [Accessed:17/06/2009], June 2009.

[4] Javier Artiles, Julio Gonzalo, and Satoshi Sekine. Weps 2 evaluation campaign: overview of the web people search clustering task. In 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference, 2009.

[5] Sekine Satoshi and Artiles Javier. Weps 2 evaluation campaign: Overview of the web people search attribute extraction task. In 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference, 2009.

[6] J.B. Lee and J.J. Kim and J.C. Park. Automatic extension of gene ontology with flexible identification of candidate terms. Bioinformatics, 22(6), 665–670, 2006

[7] A. Valarakos and G. Paliouras and V. Karkaletsis and G. Vouros, Enhancing Ontological Knowledge through Ontology Population and Enrichment, Proceedings of EKAW2004, LNAI 3257, Springer, 144-156, 2004.

[8] Atanas Kiryakov. Ontologies for Knowledge Management. In Semantic Web Technologies. Wiley, 2006

[9] The NetFlix Dataset. http://www.netflixprize.com/download - [Accessed:17/06/2009]

Document Updates

Version[8]	Date[9]	Updates and Revision History[10]	Author
0.1	26/05/2009	Draft Table of Contents	Ioannis Klapaftis
0.2	10/06/2009	Draft Deliverable 4.1 (incomplete)	Ioannis Klapaftis
0.3	15/06/2009	Draft Deliverable 4.1 (incomplete)	Ioannis Klapaftis
0.4	23/06/2009	Draft Deliverable 4.1 (incomplete)	Shailesh Pandey
0.5	25/06/2009	Draft Deliverable 4.1 (incomplete)	Ioannis Klapaftis
0.6	26/06/2009	Draft Deliverable 4.1	Ioannis Klapaftis
0.7	28/06/2009	Draft Deliverable 4.1	Ioannis Klapaftis
0.8	28/06/2009	Draft Deliverable 4.1	Shailesh Pandey
0.9	29/06/2009	Draft Deliverable 4.1	Suresh Manandhar
1.0	30/06/2009	Final Deliverable 4.1	Ioannis Klapaftis, Shailesh Pandey & Suresh Manandhar

[8] In form of “vYYYYMMDD”; Version number and edition should correspond to the actual document name conventions.

[9] In form of “DD/MM/YYYY”

[10] Attach as appendix document reviews when appropriate; describe also the current status of the document e.g. “released for internal review”, “released for comments from partners”