Advanced Language Processing Technology Applied to Digital Records: Annual Technical Status Report

Oct. 1, 2009 – Sept. 30, 2010

William Underwood
Sandra Laib
Sheila Isbell
Akilah McIntyre
Rita Gonzalez
Stan Hughes

Technical Report ITTL/CSITD 10-04

November 2010

Georgia Tech Research Institute
Information Communication Laboratory
Atlanta, Georgia

Research was sponsored by The Army Research Laboratory (ARL) and The National Archives Center for Advanced Systems and Technologies (NCAST) and was accomplished under Cooperative Agreement Number W911NF-10-2-0030. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory, NCAST, or the US Government. The U.S. government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

ABSTRACT

The overall objective of this research project is to apply computational linguistics technology to problems that arise in summarizing, accessing, reviewing and preserving electronic records of the Department of Defense and Presidential administrations. This report summarizes progress of the first year of research.

When archivists describe records or review records for possible restrictions on disclosure, they must identify actions, such as proclaiming, directing or recommending, that are carried out by the records. Linguists refer to these actions as speech acts. To support the development of a technique for automatically recognizing these speech acts, we have completed the collection of a corpus of e-records that include speech acts expressed with performative verbs. Most of these records are copies of federal or presidential e-records.

Summary translations of foreign language documents are useful to researchers when they want to know in general terms what a document is about, but full translations are not available. An English language title or scope and content note for a foreign language record is also useful to archivists or researchers who are searching for electronic foreign language records that have not been translated. Finally, indexes to the names of persons and locations in foreign language documents are also useful in locating relevant foreign language documents. To support the creation of such finding aids for Arabic language documents, we are creating Arabic language resources that can be used with the General Architecture for Text Engineering (GATE) for annotating proper nouns in Arabic documents. This research is being conducted in collaboration with the Multilingual Computing Branch of the Army Research Laboratory.

Automated file format identification and validation is a necessary feature for the ingestion of digital objects into an archive. It is also needed for determining which viewer or player is needed for a file, for recognizing archive files containing other files and for recognizing password encrypted files. To improve file format identification technology, we are extending the capabilities of the Linux file command through a greatly enhanced magic file. The GTRI file type identifier recognizes the file signatures of 850 file formats. During the current year of research additional examples of these file types have been collected and used to test the file type identifier. Tests indicate that it correctly identifies 700+ file formats. A library (database) of information about each file format identified by the file format identifier has been created that also includes file format specifications, sample files, viewers/players, converters, and archive extractors that we have collected for most of these file formats.

The National Archives (TNA) of the UK is developing a registry for file formats that includes information that can be used by DROID to identify file types. This registry is an important international resource for archives, libraries and other institutions that need to preserve and manage digital assets. The National Archives and Records Administration (NARA) is both a user of information in this registry and a partner in creating the registry. This research project has contributed file signature information to the PRONOM Registry that enhances the capability of DROID to identify file formats.

Continued access to the content of files in obsolete file formats requires migrating viewers/players for these file formats to new computer platforms, conversion of proprietary or legacy file formats into standard or current file formats, or development of media adaptors for interpreting these formats in a multivalent browser. In order to meet these challenges, the definition of file formats via formal grammars is being investigated as a means to supporting translators and viewers for file formats.

NARA's Transcontinental Persistent Archives Prototype (TPAP) is a collaborative research test bed in which the challenges inherent in preserving, protecting, and providing access to electronic records are being addressed. Previously, GTRI had a SRB (Storage Resource Broker) server as part of NARA's TPAP data grid. The SRB has been replaced by the iRODS (integrated Rule Oriented Data System) technology, a second generation data grid system providing a unified view and seamless access to distributed digital objects across a wide area network. We have created an iRODS server for integration with the TPAP data grid. It is anticipated that it will go online in January 2011.

TABLE OF CONTENTS

1 Introduction

The overall objective of this research project is to apply computational linguistics technology to problems that arise in summarizing, accessing, reviewing and preserving electronic records of the Department of Defense and Presidential administrations. Among issues and problems to be addressed in cooperation with ARL are technology areas responsive to the Army's missions, including automated discourse analysis, automatic summarization, content-based information retrieval, rule-based reasoning, and network-centric/distributed information systems technology. This research is being conducted in collaboration with the Multilingual Computing Branch of the Army Research Laboratory and the NCAST Research Laboratory of the National Archives and Records Administration

In the following sections, progress on the research tasks of the first year is described.

2 Recognizing the Speech Acts Performed by E-records

All written records of human activity involve actions expressed by the records. Archivists identify these acts when they describe records and when they review records for possible restrictions on disclosure.

Underwood [2008] analyzed 120 Presidential records to determine the occurrence of explicit, implicit and indirect speech acts occurring in the records, as well as the use of textual structure to indicate speech acts. It was shown that the speech acts identified in the records could be used to construct a description for the records that indicated the primary action of the records.

A method for automatically recognizing these acts was formulated. The first step of this method is to identify explicit speech acts in records that are expressed in performative sentences with performative verbs.

Vanderveken has identified and defined 271 performative verbs. However, he provided only a few examples of performative sentences for these performative verbs. The Public Papers of the Presidents [1991-2005] were searched for examples of performative sentences using these performative verbs. Examples were found for 152 performative verbs defined by Vanderveken. Examples of performative sentences were also found for 49 additional performative verbs not defined by Vanderveken [Underwood 2009b].

Examples were sought for the remaining 119 performative verbs defined by Vanderveken for which examples had not been found in Presidential Papers. The archives of the American Presidency Project[1] were searched for texts containing performative uses of performative verbs. Google was also used to search the internet at large for texts containing performative uses of performative verbs. Examples of performative sentences for all 119 of these performative verbs have been found [Underwood 2010a]. These

[1] www.presidency.ucsb.edu/index.php

examples of performative sentences are being analyzed to define sentence patterns in JAPE rules that can be used to recognize performative sentences.

3 Automatic Content Extraction from Arabic Language Documents

An English summary translation of a foreign language document is a broad overview of the content and conclusions of a document in summary form. Summary translations are useful to researchers when they want to know in general terms what a document is about, and when a full translation would be too time consuming or expensive to produce. An English language title or scope and content note for a foreign language record (or file unit or series of such records) would also be useful to archivists or researchers who are searching for electronic foreign language records that have not been translated.

This research task is to collaborate with Computational Linguists/Computer Scientists at ARL, first, in developing a method to automatically annotate person's names, location names, names of governments, and dates in documents in the Arabic language. Second, these annotations will be used to construct indexes to the documents. Third, these annotations will be used to construct English language titles and scope and content notes for the documents. The Multilingual Computing Branch, Information Sciences Division of ARL is exploring the use of Kepler, a scientific workflow system, and Moses, a statistical machine translation toolkit, for supporting semi-automated translation of languages such as Arabic and Pashto.

Our first step is to develop a high performance method for named-entity recognition in Arabic records. We familiarized ourselves with the Arabic plug-in provided with GATE which contains a vanilla application for Arabic named entity recognition. The application contains processing resources for tokenization, wordlist lookup, and orthographic co-reference. There is also an application to collect new Arabic word lists from training data.

Figure 1 shows file names, descriptions, number of entries and the type:subtype of annotations for the 25 lists making up the Arabic lexical resources provided with the vanilla version of the GATE Arabic plug-in. The Arabic wordlists are used for initial annotation of the words in an Arabic text document. The Java Annotation Pattern Engine (JAPE) rules provided with the plug-in only convert the initial annotations to final annotations. They do not combine elements such as given names and family (clan) names or city and country names. Furthermore there is no Arabic lexicon provided with the plug-in.

File Name	**Description**	**Entries**	**Type: Subtype**
ordinal	For example, *alawul*, "the first."	55	number:ordinal
city	Names of cities in Arabic countries, for example, *trabulus*, "Tripoli".	85	location:city
city_world	Arabic names for other cities of the world, for example, *brleen*, "Berlin".	127	location:city
conjunction	Arabic conjunctions, for example, *uu*, "and."	3	conjunction

File Name	Description	Entries	Type: Subtype
country	Names for countries and emirates in North Africa and Middle East, for example, *mesr*, “Egypt” and *abu dhabi*, “Abu Dabi.”	22	location:country
country_world	Arabic names for other countries in the world, for example, *zaambyaa*, “Zambia.”	171	location:country
currency	Names of major currencies in the world, for example, *dulaar*, “Dollar.”	17	money_unit
date_key	Arabic words denoting types of calendar, for example, *hijri*, literally “the emigration of the Prophet Muhammad from Mecca to Medina,” used to denote the Islamic or Lunar Calendar.	4	date_key
days	Arabic names for days of the week, for example, *al-khmees*, “Thursday.”	7	date:day
facility	Names of specific places of religious, political, and social significance, for example, *al-masjid al-aqsa* “Al-Aqsa Mosque.” Also, Arabic words for non-specific places, for example, *kanees*, “synagogue.”	35	facility
female_names	Arabic first names for women which can have meaning too, for example, *aya,* means “verse.”	708	person:female
location_other	Arabic names for bodies of water, islands, and continents, for example, *albaaseefeek mheet*, “Pacific Ocean,” and *seeshel hzr*, “Seychelles Islands,” and *aseeaa qaarat*, “Asia Continent.”	76	location:other
male_names	Arabic first names for men which can have meaning too, for example, *asad*, means “lion.”	829	person:male
months	Arabic names for Lunar and Gregorian calendars. For example, *sfr*, “safar-second month of lunar calendar, or *deesmbr*, “December.”	35	date:month
monuments	Names of notable landmarks outside the Arab World, for example, *borj eyfel*, “Eiffel Tower.”	3	facility:monument
mountains	Arabic names of mountains in the world, for example, *jabal al-alb*, “Alp Mountains.”	30	location:mountain
oceans_seas_islands	See above word list, “location_other.”	76	location:other
ordinals	For example, *alfee*, “thousand.”	55	number:ordinal
organizations	Names and acronyms of international organizations, government agencies, media networks: for example, *al-bee bee see*, “the BBC,” or *al-amen al-mutahadat*, “United Nations, or *musaad*, “Mossad.”	96	organisation
percent	Arabic word meaning “percent,” *bamaat*.	3	percent
places	Arabic word for public places in a city or town, for example, *wuuq*, “market” or *mustashfa* “hospital.”	52	location:other
rivers	Arabic names for rivers in the world, for example, *naher al-raayen*, “Rhine River”	40	location:river
surnames	Arabic family names, for example, *al-shmree*.	197	person:surname
time	Arabic words describing time, for example, *saa3,*”hour” or *shaher*, “month.”	43	time
titles	Names ascribed to certain positions or professions, for example, *dhaabt*, “officer” or *amir*, “prince.”	272	person:title

Figure 1. Arabic Wordlists Provided with GATE Arabic Plug-in

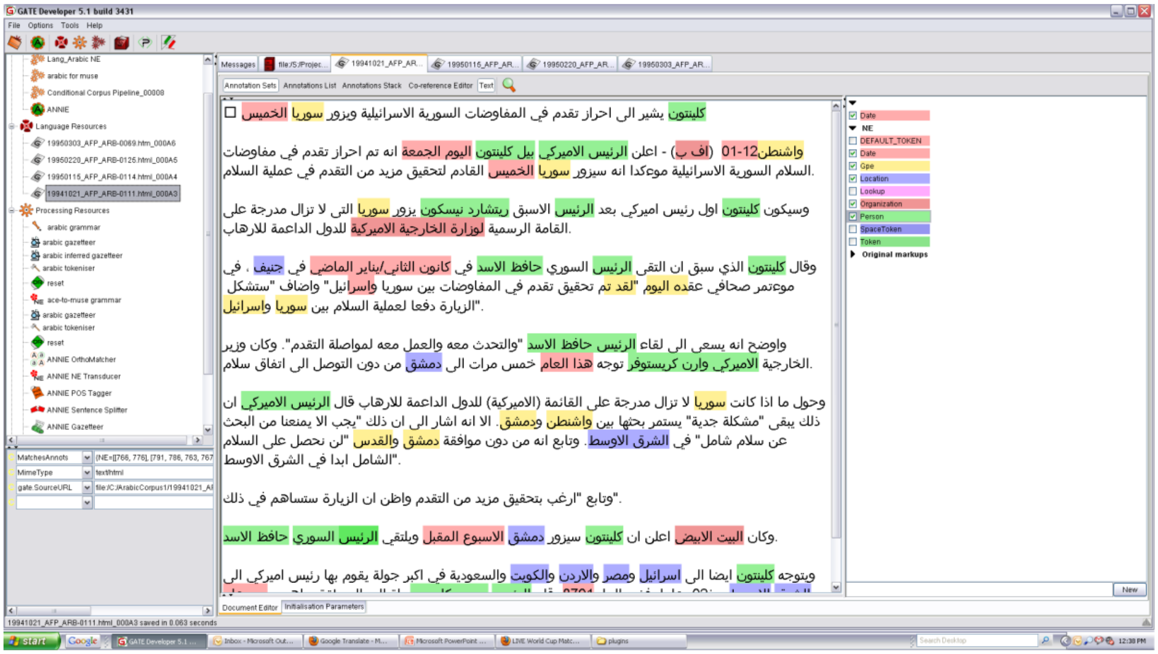
The GATE 5.1 distribution contains a second set of Arabic lexical resources. They are 17 lists extracted from the Automatic Content Extraction (ACE) 3 v1.2 training set of 119 files from broadcast news and news wires. They are in a folder titled gazetteers-inferred. The lists are described in Figure 2. Note that person names are full names, not separate lists for given names and family names. The terms for money are the exact amounts found in the corpus, not separate lists for amounts and units. The lists will work very well for demonstration purposes on the documents from which they were extracted, but not well on Arabic documents in general. While there are JAPE rules provided with these wordlists, they only promote the initial annotations to final annotations. They do not perform the more general annotation of combinations of more elemental terms.

FILE NAME	**DESCRIPTION**	**ENTRIES**	**TYPE:SUBTYPE**
CARDINAL	Numerals that consist of whole numbers, “decimals” which are denoted with commas as in the European convention, as well as spelled out numbers. **Ex**: 00631, 88, *alf 48* “thousand 48,” *5,1 mleeuun* “5.1 million,” *althltheen* “thirty.”	316	Cardinal:inferred
DATE	Denotes instances of time, date, and duration. **Ex**: *ashreen aamaa* “years twenty,” *layl* “night,” *aljmaa* “Friday,” *yuumeen* “two days,” *thlatha aw arbaa ayaam* “three or four days,” *asbuueen* “two weeks,” *ayluul/sbtmber* “September,” and *alashr alakheerat* “recent months.”	916	Date:inferred
FAC	References to unspecific man-made structures such as sports venues, places of worship, and some types of infrastructure. Also, there are some proper names of specific locations. **Ex**: *tshrnuubeel* “Chernobyl” *astaad* “stadium,” *jsr alemeer mhmd* “Prince Mohammed Bridge,” *ghrft* “room,” *alatbaat alsheeat almqdsat* “Shiite holy sites,” *albayt alabyd* “White House,” and *mstuutnaat* “settlements.”	92	Fac:inferred
FAC_DESC	Descriptors of types of structures. **Ex**: *shaara* “street,” *almrakz* “centers,” *nfq* “tunnel,” *mbna* “building,” *mktb* “tunnel,” *kaatdraat* “cathedral,” *almqr alaam* “headquarters” *daar* “house.”	107	Fac_desc:inferred
GPE	General references to cities, districts, and provinces. Specific references to proper names of cities, countries, and regions. **Ex**: *aasmat* “capital,” *iraakee* “Iraqi,” *umaan* “Oman,” *zeembaabyee* “Zimbabwe,” *seednee* “Sydney,” *aaseea* “Asia.”	362	Gpe:inferred
GPE_DESC	Varied references to cities, states, countries, and regions. **Ex**: *mdeena* “city,” *alwaalyaat* “states,” *ljmhuureeat* “of the republic of,” *wtn* “homeland,” *arkhbeel* “archipelago.”	89	Gpe_desc:inferred
LOC	General and specific references to territories, boundaries, natural formations as well as locations in the Middle East. **Ex**: *taaeemz skueer* “Times Square,” *ardh* “land,” *ardh msr* “the land of Egypt,” *hduud* “border,” *areehaa* “Jericho,” *shmaal* “north.”	60	Location:inferred
MONEY	Instances of U.S. monetary values. Symbols of monetary unit are not included in this list. **Ex**: *alf duular 52* “52 thousand dollars,” mlyaar duular 5,01 “$5.1 billion,” *mleeuum duular 52* “$52 million,” *mlayeen men alduularaat* “millions of dollars,” *khmseen sntaa* “fifty cents.”	88	Money:inferred

FILE NAME	DESCRIPTION	ENTRIES	TYPE:SUBTYPE
NATIONALITY	Instances of adjectival form of countries, religions, and political ideologies. **Ex**: *alameerkeea* “American” (feminine), *alameerkee* “American” (masculine), *albraazeelyeen* “Brazilian,” *alflsteenyeen* “Palestinian,”*alyhuud* “Jews,” *alnaazyeen aljdd* “Neo-Nazis,” *almslmeen* “Muslims,” *ghrbyeen* “western.”	459	Nationality:inferred
ORDINAL	List of ordinal numbers. **Ex**: *alawla* “first,” *khaamsat* “fifth,” *al 23* “the 23,” *althaamn w alashreen* “twenty-eighth.”	62	Ordinal:inferred
ORG	Names and/or acronyms of companies, political parties, international organizations, and government agencies. **Ex**: *saamsuungghr* “samsung,” *see bee as* “CBS,” *hzb* “party,” *asuusheetdh brs* “Associated Press,” *hzb alleekuud alymeenee alisraeeli* “Israeli right-wing Likud party,” *suut amreekaa* “Voice of America,” *alkhtuut aljuuyat almlkeeat alurdneeyat* “Royal Jordanian Airlines.”	185	Organization:inferred
PER	General words referring to people, men, women, child, and types of professions. Some names of fictional characters as well as proper names. **Ex**: *seelfeeuu brlskuunee* “Silvio Berlusconi,” *hshdh* “crowd,” *dhaayat* “advocate,” *akhee* “my brother,” *rbna* “our Lord,” *bruubn kuuk* “robin cook.”	611	Person:inferred
PER_DESC	References to a person or group of people. Also job titles included in this list. **Ex**: *almshrdeen* “displaced,” *almsuureen* “photographers,” *raees* “president,” *rjl alaamaan* “businessman,” *shaahdh* “witness.”	874	Per_desc:inferred
PERCENT		48	Percent:inferred
PRODUCT_DESC	Modes of transportation, artillery, weapons, and machinery. **Ex**: *baas* “bus,” *taayaarat* “plane,” *buaakhr* “ship,” *jraafat* “bulldozer,” *alseef* “swords.”	132	Product_desc:inferred
QUANTITY	Instances of measurements. **Distance**: *thlatheen klm* “thirty kilometers.” **Volume**: *alf 005 brmeel* “5000 barrels.” **Area**: *06 mtraa mrbaa* “6 square meters.” **Temperature**: *2,6 drjaat* “2.6 degrees.” **Speed**: *thlatheen mtra fee althaaneeat* “thirty meters per second.” **Energy**: *mleeuun keeluuat 006* “6 million kilowatts.” **Weight**: *khmsat mlayeen tn* “5 million tonnes.”	113	Quantity:inferred
SUBSTANCE	References to different types of **drinks**: *alnbeedz* “wine,” *albeerat* “beer.” **Food**: *hluuyatt* “candy,” *albeed* “eggs.” **Other**: *alnft* “oil,” *alhmdh alreebee alnuuwee* “DNA,” *alghraz almseel lldmuuaa* “tear gas” *aldhm* “blood.”	71	Substance:inferred

Figure 2. Arabic Wordlists Inferred from ACE Training Set

Figure 3 shows the Graphical User Interface to GATE and the results of automatically annotating the named-entities (location, person, organization names and dates, times, percent and currency) in an Arabic document. The lexical resources used were the inferred gazetteers (wordlists) and the Arabic language document was one of the documents in the training set. Consequently the annotation is fairly precise.



MatchesAnnots	{NE=[[766, 776], [791, 786, 763, 767
MimeType	text/html
gate.SourceURL	file:/C:/ArabicCorpus1/19941021_AI

19941021_AFP_ARB-0111.html_000A3 saved in 0.063 seconds

☑ Date

NE
- ☐ DEFAULT_TOKEN
- ☑ Date
- ☑ Gpe
- ☑ Location
- ☐ Lookup
- ☑ Organization
- ☑ Person
- ☐ SpaceToken
- ☐ Token

Original markups

□ كلينتون يشير الى احراز تقدم في المفاوضات السورية الاسرائيلية ويزور سوريا الخميس

واشنطن01-12 (اف ب) - اعلن الرئيس الاميركي بيل كلينتون اليوم الجمعة انه تم احراز تقدم في مفاوضات
السلام السورية الاسرائيلية موءكدا انه سيزور سوريا الخميس القادم لتحقيق مزيد من التقدم في عملية السلام.

وسيكون كلينتون اول رئيس اميركي بعد الرئيس الاسبق ريتشارد نيسكون يزور سوريا التى لا تزال مدرجة على
القامة الرسمية لوزارة الخارجية الاميركية للدول الداعمة للارهاب.

وقال كلينتون الذي سبق ان التقى الرئيس السوري حافظ الاسد في كانون الثاني/يناير الماضي في جنيف ، في
موءتمر صحافي عقده اليوم "لقد تم تحقيق تقدم في المفاوضات بين سوريا واسرائيل" واضاف "ستشكل
الزيارة دفعا لعملية السلام بين سوريا واسرائيل".

واوضح انه يسعى الى لقاء الرئيس حافظ الاسد "والتحدث معه والعمل معه لمواصلة التقدم". وكان وزير
الخارجية الاميركي وارن كريستوفر توجه هذا العام خمس مرات الى دمشق من دون التوصل الى اتفاق سلام.

وحول ما اذا كانت سوريا لا تزال مدرجة على القائمة (الاميركية) للدول الداعمة للارهاب قال الرئيس الاميركي ان
ذلك يبقى "مشكلة جدية" يستمر بحثها بين واشنطن ودمشق. الا انه اشار الى ان ذلك "يجب الا يمنعنا من البحث
عن سلام شامل" في الشرق الاوسط. وتابع انه من دون موافقة دمشق والقدس "لن نحصل على السلام
الشامل ابدا في الشرق الاوسط".

وتابع "ارغب بتحقيق مزيد من التقدم واظن ان الزيارة ستساهم في ذلك".

وكان البيت الابيض اعلن ان كلينتون سيزور دمشق الاسبوع المقبل ويلتقي الرئيس السوري حافظ الاسد

ويتوجه كلينتون ايضا الى اسرائيل ومصر والاردن والكويت والسعودية في اكبر جولة يقوم بها رئيس اميركي الى

Figure 3. Automatic annotation of the named entities in an Arabic document

To initially evaluate the performance of GATE's Arabic named entity recognition plug-in, a corpus of 50 Arabic news wires, newspaper and magazine articles was collected.[2] The named entities in a copy of the corpus were manually annotated. This manually correctly annotated copy of the corpus is referred to as the key.

The corpus of 50 documents was automatically annotated using the vanilla copy of the Arabic wordlists provided with the GATE 5.1 distribution. This automatic annotation of the corpus was compared with the key to determine the performance of the current Arabic lexical and processing resources. The performance was so poor as to not be reported. The performance was poor due to so few terms in the wordlists and no JAPE rules for performing functions such as combining annotated given names and family names into full names.

We have begun to create additional and larger Arabic wordlists. We are also creating JAPE rules that create annotations for terms made up of more primitive Arabic terms. We are also creating an Arabic lexicon that is needed for annotating Arabic parts of speech [Gonzalez et al 2010]. When these lexical resources are more fully developed we will experimentally test them on an Arabic Corpus.

[2] Corpus of Contemporary Arabic (CCA) www.comp.leeds.ac.uk/eric/latifa/research.htm

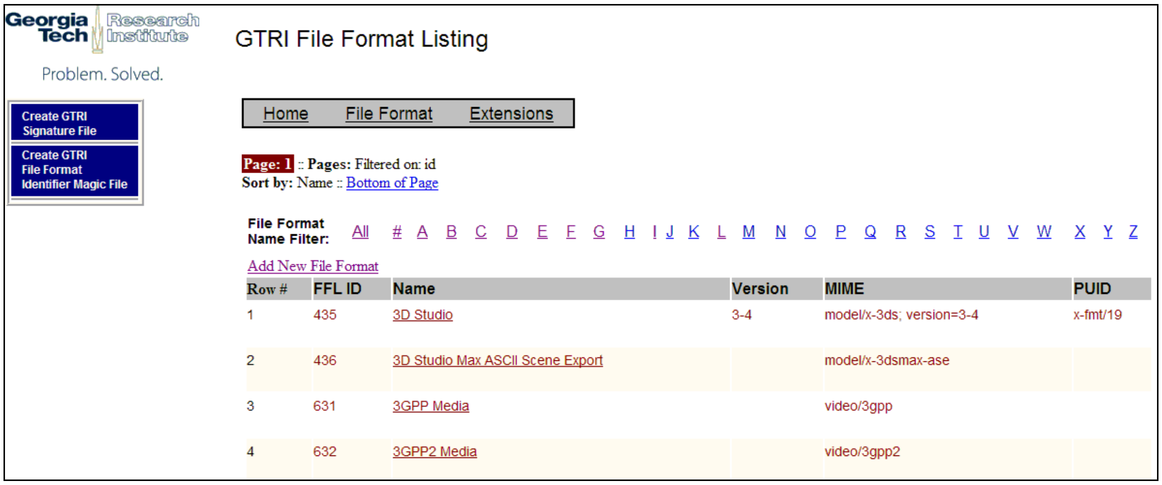
4 Automatic File Format Identification

Automated file format identification and validation is a necessary feature for the ingestion of digital objects into an archive. Extraction of file metadata is also required.

The Linux *file* command and *magic* file are one of the most promising technologies for file type identification and metadata extraction. We are extending its capabilities through a greatly enhanced magic file of file signatures and a File Format Library [Underwood 2009a].

4.1 GTRI File Format Library

The File Format Library is implemented using JAVA and MYSQL and is Web-enabled. The library includes information for the 850 file formats currently identified by the GTRI File Type identifier. Figure 4 shows part of the interface to the File Format Library for browsing file formats in the Library by file format name, MIME-type and PRONOM Unique identifier (PUID).


Georgia Tech Research Institute
Problem. Solved.
Create GTRI Signature File
Create GTRI File Format Identifier Magic File
Home
File Format
Extensions
File Format Name Filter:
All # A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
Add New File Format

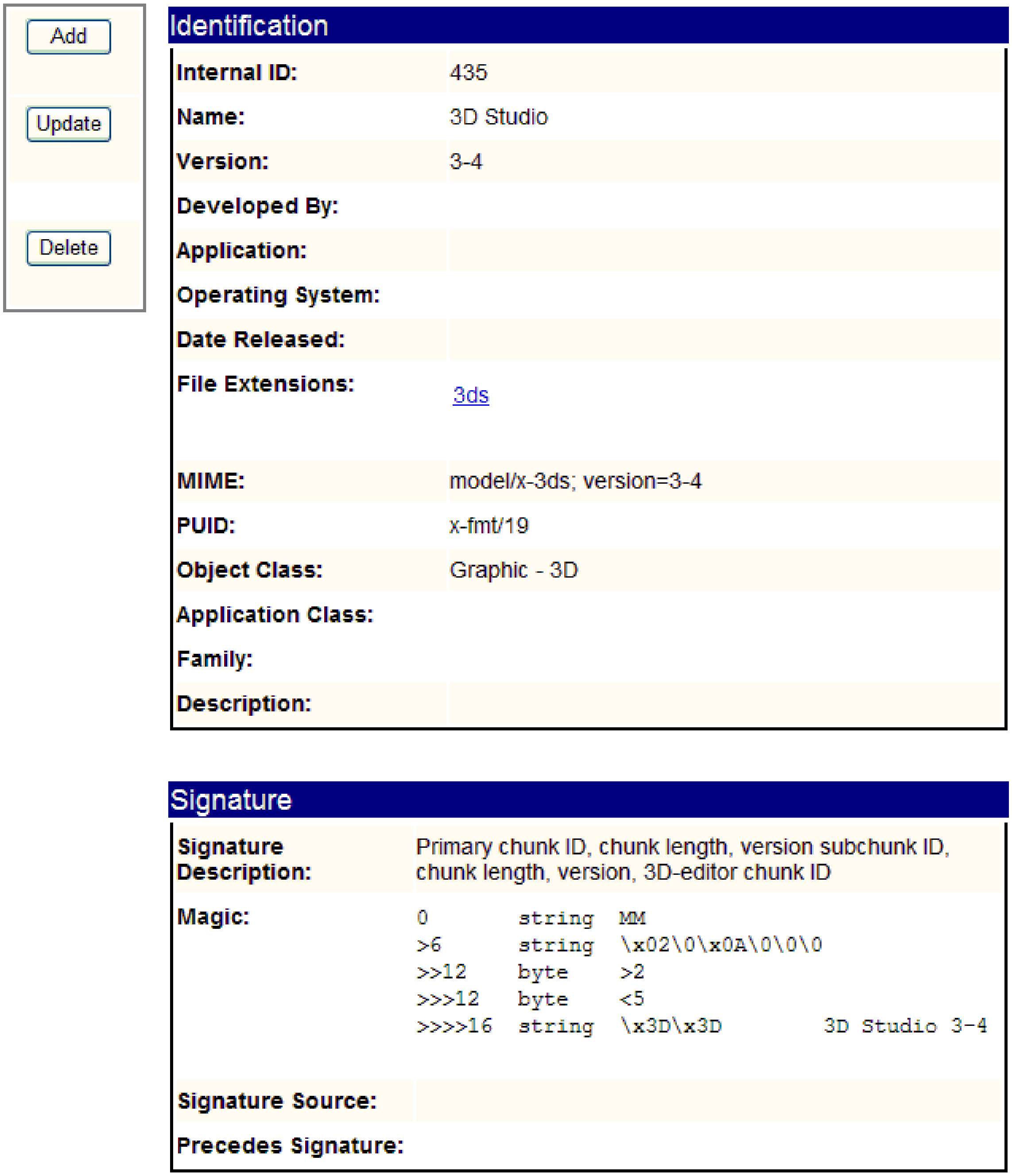
GTRI File Format Listing

Page: 1 :: **Pages:** Filtered on: id
Sort by: Name :: Bottom of Page

Row #	FFL ID	Name	Version	MIME	PUID
1	435	3D Studio	3-4	model/x-3ds; version=3-4	x-fmt/19
2	436	3D Studio Max ASCII Scene Export		model/x-3dsmax-ase	
3	631	3GPP Media		video/3gpp	
4	632	3GPP2 Media		video/3gpp2	

Figure 4. Browsing the File Format library

Figure 5 shows the interface for adding file format identification and signature information to the Library as well as editing this information and deleting file formats.



Identification

Internal ID:	435
Name:	3D Studio
Version:	3-4
Developed By:	
Application:	
Operating System:	
Date Released:	
File Extensions:	3ds
MIME:	model/x-3ds; version=3-4
PUID:	x-fmt/19
Object Class:	Graphic - 3D
Application Class:	
Family:	
Description:	

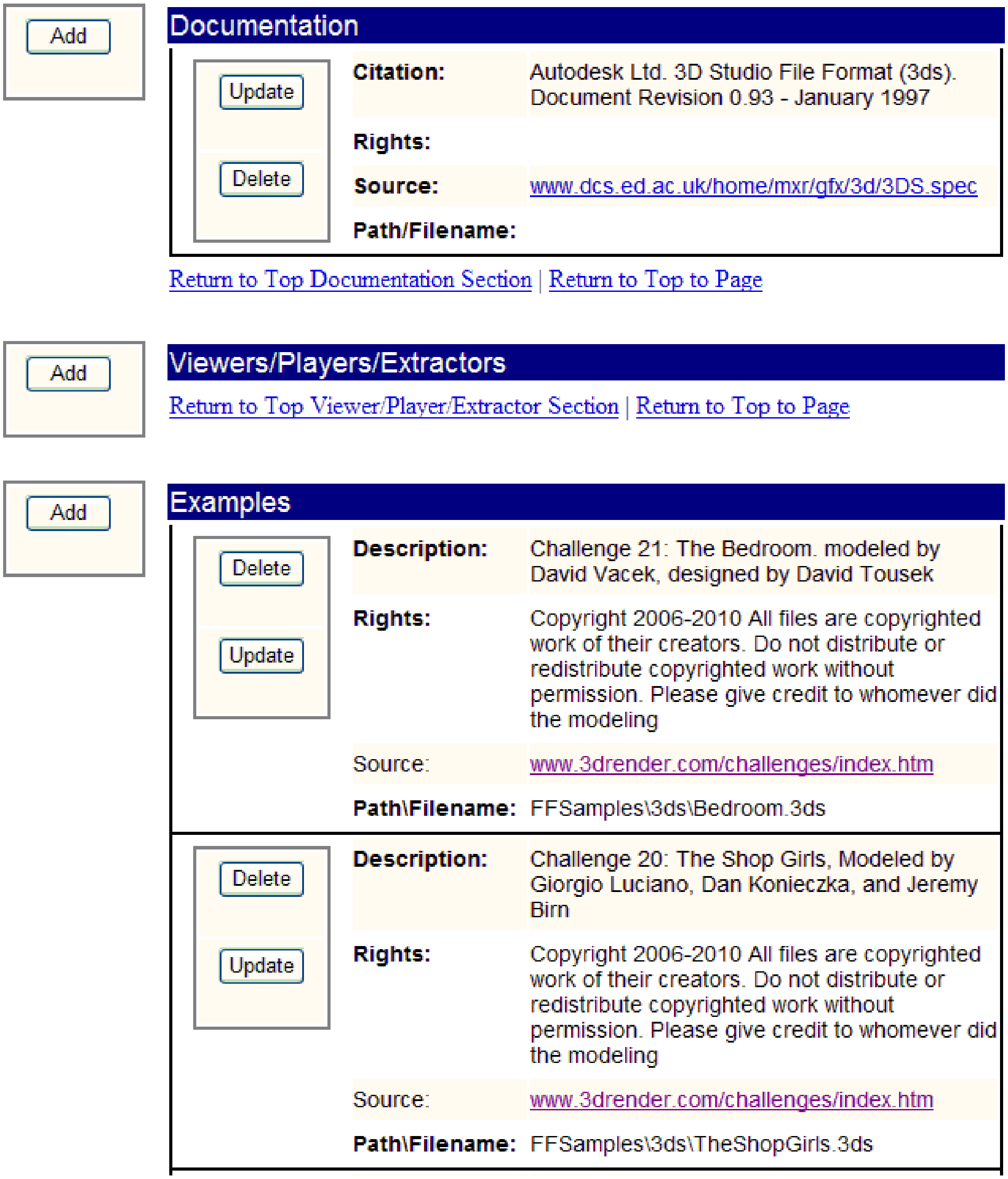
Signature

Signature Description:	Primary chunk ID, chunk length, version subchunk ID, chunk length, version, 3D-editor chunk ID
Magic:	see below
Signature Source:	
Precedes Signature:	

```
0       string  MM
>6      string  \x02\0\x0A\0\0\0
>>12    byte    >2
>>>12   byte    <5
>>>>16  string  \x3D\x3D        3D Studio 3-4
```

Figure 5. File format Identification and Signature Information

Figure 6 shows information on documentation of the files formats and sample files using the format.



Documentation

Citation: Autodesk Ltd. 3D Studio File Format (3ds). Document Revision 0.93 - January 1997

Rights:

Source: www.dcs.ed.ac.uk/home/mxr/gfx/3d/3DS.spec

Path/Filename:

Return to Top Documentation Section | Return to Top to Page

Viewers/Players/Extractors

Return to Top Viewer/Player/Extractor Section | Return to Top to Page

Examples

Description: Challenge 21: The Bedroom. modeled by David Vacek, designed by David Tousek

Rights: Copyright 2006-2010 All files are copyrighted work of their creators. Do not distribute or redistribute copyrighted work without permission. Please give credit to whomever did the modeling

Source: www.3drender.com/challenges/index.htm

Path\Filename: FFSamples\3ds\Bedroom.3ds

Description: Challenge 20: The Shop Girls, Modeled by Giorgio Luciano, Dan Konieczka, and Jeremy Birn

Rights: Copyright 2006-2010 All files are copyrighted work of their creators. Do not distribute or redistribute copyrighted work without permission. Please give credit to whomever did the modeling

Source: www.3drender.com/challenges/index.htm

Path\Filename: FFSamples\3ds\TheShopGirls.3ds

Figure 6. File format Documentation and Sample Files

4.2 GTRI File Format Identifier

Significant progress has also been made in the definition of the *gtri-magic* file. The original GTRI File Type Identifier used the 4.21 version of the *file* command. The *gtri-magic* file is being migrated to the newest version (5.04) of the *file* command. However, changes in the 5.04 release require additional modifications and testing of the *gtri-magic* file. The upgrade and testing of the *gtri-magic* file with the new 5.04 release of the *file* command is scheduled to be completed in January 2011.

4.3 Contributions to PRONOM Registry

Fifty of the signature tests for file formats in the GTRI File Type Identifier were converted to the regular expression notation used in the PRONOM registry for defining tests for file format signatures. These were shared with The National Archives (TNA) PRONOM Program. Also shared was the rationale for each of the signatures, links to specifications for the file formats and samples of files for each of the file formats. The staff of the PRONOM program has tested these file signatures and have incorporated them into the next release of the DROID signature file.

4.4 Collaboration with the NCAST Lab

Mark Conrad and student interns at the NCAST Research Laboratory have been creating a repository of digital records created by Federal agencies. We provided this laboratory with a table of 240 file format names and file name extensions for file formats with 3D data content. Examples and specifications are needed for these file formats to develop file signature tests and to test the performance of the GTRI File Type Identifier. Examples were needed that were produced by US Government agencies because many of these are in the public domain and thus can be shared with other researchers. The student interns searched their existing repository and the Internet and located file format specifications for about 30 of these formats and about 5000 examples for about 40 of the file formats. We subsequently provided the laboratory with a complete list of the more than 850 file formats for which signature tests have been developed but for which additional examples are needed. We developed about 60 additional signature tests for file formats with 3D data content based on the contributions of the NCAST lab.

5 Grammars and Parsers for File Formats

This research task is to investigate whether formal grammars can be extended to define the structure and semantics of file formats. One of the most general classifications of file formats is that of file formats into text files and binary files. Text files are computer files that are structured as a sequence of lines of text characters. Binary files are structured as a sequence of bytes that include data types other than text characters such as short and long integers, pointers to relative addresses, representations of image gray-levels or colors, and representations of sounds.

The file formats in the File Format Library (currently 850 file formats) have been classified into 235 Text File Formats and 615 Binary File Formats. During the first year of research, we focused on identifying specifications for text file formats that were in terms of formal grammars or in creating grammars for those not specified in terms of grammars. [Underwood 2010b]

Our initial findings are that plain text files for ASCII encodings, DOS Extended ASCII, the ISO-8869 family of character sets, EBCDIC, Microsoft Code Page 1252, MacRoman, and the Unicode character sets can all be defined using regular expressions and thus are also definable as regular grammars. AWK, Perl, UNIX shell scripts, and DOS batch files can all be defined with context-free grammars. Source code files for programming languages are uniformly defined using context-free grammars. These include the programming languages FORTRAN, COBOL, C, C++, LISP/Scheme, Pascal and PYTHON.

All XML documents and all markup languages based on SGML or XML are definable with context-free grammars. This is because Document Type Definitions (DTDs) are extended context-free grammars.

We are still examining the file formats for formatted text files to determine whether they can all be defined with context-free grammars. These formats include Adobe PostScript, troff input text, Framemaker Book, Framemaker interchange format, TeX documents, Ami Professional documents, Applix Word documents, Chiwriter documents, Mathematica Notebooks, Pocket Word documents, RTF Documents, and XyWrite documents.

There are several spreadsheets and spreadsheet interchange formats that have text formats, for instance, Applix spreadsheet, Data Interchange Format, Comma-separated values, Tab-separated values, and SYLK files. There are 17 Printed Circuit Board Netlists (CAD files) that have textual formats.

The really interesting textual file formats are those for bitmap graphics, and 2D- and 3D-vector graphics. The file type identifier currently identifiers 4 families of bitmapped graphics file formats, 10 families of 2D-vector graphics file formats, and 32 families of 3D-file formats that have text file formats. Some of the file format specifications of these file formats are in terms of context-free grammars that specify a geometric language for describing 2-D or 3-D geometric objects.

In the coming year we will complete the analysis of the text file formats and demonstrate how a parser and interpreter can be used for validation and display of some of the objects encoded in those formats. Then we will begin the analysis of the binary file formats.

The significance of success in this research task is that if most file formats can be defined as extensions of RGs, CFGs and CSGs, then only three parsing/translation algorithms are needed for verifying that a file conforms to a particular format. Similarly, only three translation algorithms are needed for conversion of legacy file formats to current or standard formats. Finally, only three algorithms are needed for viewer/players for most file formats. This would increase the likelihood of preserving and making available digital records into the indefinite future.

6 Services for the Transcontinental Persistent Archive Prototype

Dr. Underwood and Akilah McIntyre attended the iRODS User Meeting 2010, March 24-26 at the RENCI Center, Chapel Hill, North Carolina.[3] iRODS stands for integrated Rule-Oriented Data System. It is a technological framework for implementing DataGrids or access to distributed heterogeneous database systems. The meeting included tutorials on existing micro-services and how to implement micro-services and icommands and irules using micro-services. Micro-services are C-functions that may be calling iRods' client or server routines. Irules are used for implementing access, security and replication policies.

One of our research subtasks is to implement an *ifile* command using the NetBSD Fine Free File Command known as *file*.[4] The *file* command ships with every free operating system (OpenBSD, Linux, NetBSD, FreeBSD, etc.) and has been ported to most systems (OS/2, DOS, MS Windows, etc.). The functionality will be implemented with calls to the iRODS micros-services using the *file* command's *libmagic* library. Having implemented the *ifile* command, the current GTRI File Type Identifier GUI can be implemented as an iRODS client. There is an iRODS Windows client for data grids called iRODS Explorer. It may be desirable to integrate the File Type identifier into the iRODS Explorer as a plug-in.

Previously, GTRI had a SRB (Storage Resource Broker) server as part of NARA's Transcontinental Persistent Archive Prototype (TPAP) data grid. The SRB has been replaced by the iRODS technology. Stan Hughes and Akilah McIntyre created the following iRODS servers for integration and development: an internal Ubuntu 10.04 desktop system for software development, an internal Ubuntu 10.04 server for testing, and an internet facing Red Hat server for integration with the TPAP data grid. Connection to the TPAP data grid requires Internet2 network connectivity using the IPV6 protocol. The GTRI/ICL IPV6 roll out is scheduled to be completed by January 15, 2011. At this time, the internet facing server will go online.

The current software development of the *ifile* command uses the newest version (5.04) of the *file* command. Stan Hughes started the software development of the *ifile* command with the creation of an *ifile* shell command (iRODS/clients/icommands/src/ifile.c), a file utility interface (iRODS/lib/core/src/fileUtil.h) and stubs for the utility (iRODS/lib/core/src/fileUtil.c). The next step is to create 'C' routines that will integrate with the iRODS micros-services. These routines will call the *file* command's *libmagic* library to support the functionality defined in the file utility interface.

[3] www.irods.org/index.php/iRODS_User_Meetings
[4] www.darwinsys.com/file

7 Project Management, Communication and Dissemination

During December 2009, Bill Underwood attended the International Digital Curation Conference in London and presented a peer-reviewed paper reporting results of prior research sponsored by ARL and NARA [Underwood and Laib 2009; Underwood 2009c].

During January 2010, Dr. Underwood participated in a teleconference with NARA personnel on the capabilities of the File Type Identifier.

During late March 2010, the ARL COR visited the Project Team in Atlanta. Progress on the first year research tasks was presented. The GATE interface for Arabic named-entity recognition, the File Type Identifier and the File Type Library were demonstrated.

During June 2010, Dr. Underwood was at Archives II in College Park. With NARA staff, he discussed research results that might have potential application to NARA Records Management. He also met with staff of the Army Research Laboratory to discuss cooperative research in applying information extraction, machine translation, and automatic summarization to the indexing and archival description Army and US Federal records and other documents in foreign languages. He also made a presentation entitled "File Format Identification: Progress Report" and demonstrated the GTRI File Format Library and File Type identifier. This presentation was videotaped and edited and is now web accessible via YouTube video stream www.youtube.com/watch?v=dVMs5YnZ0HU.

On December 6, 2010, the PRONOM Program of The National Archives (UK) will release a new version of the file signature file for DROID. This new release includes more than 50 signature tests that were contributed by GTRI and that were developed under ARL and NCAST sponsorship of this project.

References

[Gonzalez et al 2010] R. Gonzalez, S. Isbell and W. Underwood. Named-Entity Recognition in Arabic Documents. Working Paper ITTL/ITDSD 10-02, Information Technology and Decision Support Division, ITTL, GTRI, in preparation

[Underwood 2008] W. Underwood. Recognizing Speech Acts in Presidential E-records. PERPOS TR ITTL/CSITD 08-03, Georgia Tech Research Institute, October 2008.

[Underwood 2009a] W. Underwood. Extensions of the UNIX File Command and Magic File for File Type Identification. Technical Report ITTL/CSITD 09-02, September 2009

[Underwood 2009b] W. Underwood. Examples of Performative Sentences in Presidential Records. Working Paper ITTL/CSITD 09-01, September 2009

[Underwood 2009c] W. Underwood. Grammar-based Recognition of Documentary Form and Metadata Extraction. *International Journal of Digital Curation*. Issue 1, Vol. 5, 2010. www.ijdc.net/index.php/ijdc/issue/current

[Underwood 2010a] W. Underwood. Additional Examples of Performative Sentences. Technical Report ITTL/ITDSD 10-01, Information Technology and Decision Support Division, ITTL, GTRI, December 2010.

[Underwood 2010b] W. Underwood. Grammars for Text File Formats. Working Paper ITTL/ITDSD 10-03, Information Technology and Decision Support Division, ITTL, GTRI, in preparation.

[Underwood and Laib 2009] W. Underwood and S. Laib. A Grammar and Parser for Recognition of Documentary Forms and Metadata Extraction. TR 09-06, Information Technology and Telecommunications Laboratory, Georgia Tech Research Institute, Atlanta, Georgia, September 2009 [Revised September 2010]