Integrated Feasibility Experiment for Bio-Security: IFE-Bio A TIDES Demonstration

Lynette Hirschman, Kris Concepcion, Laurie Damianos, David Day, John Delmore, Lisa Ferro, John Griffith, John Henderson, Jeff Kurtz, Inderjeet Mani, Scott Mardis, Tom McEntee, Keith Miller, Beverly Nunan, Jay Ponte, Florence Reeder, Ben Wellner, George Wilson, Alex Yeh
The MITRE Corporation
Bedford, Massachusetts, USA and
McLean, Virginia, USA
781-271-7789
lynette@mitre.org

ABSTRACT

As part of MITRE's work under the DARPA TIDES (Translingual Information Detection, Extraction and Summarization) program, we are preparing a series of demonstrations to showcase the TIDES Integrated Feasibility Experiment on Bio-Security (IFE-Bio). The current demonstration illustrates some of the resources that can be made available to analysts tasked with monitoring infectious disease outbreaks and other biological threats.

Keywords

Translation, information extraction, summarization, topic detection and tracking, system integration.

1. INTRODUCTION

The long-term goal of TIDES is to provide delivery of information on demand in real-time from live on-line sources. For IFE-Bio, the resources made available to the analyst include e-mail, news groups, digital library resources, and eventually (in later versions), topic-specific segments from broadcast news. Because of the emphasis on global monitoring, there is a need to process incoming information in multiple languages. The system must deliver the appropriate information content in the appropriate form and in the appropriate language (taken for now to be English). This means that the IFE-Bio system will have to deliver news stories, clusters of relevant documents, threaded discussions, alerts on new events, tables, summaries (particularly over document collections), answers to questions, graphs and geo-spatial temporal displays of information.

The demonstration system for the Human Language Technology Conference in March 2001 represents an early stage of the full IFE-Bio system, with an emphasis on end-to-end processing. Future demonstrations will make use of MITRE's Catalyst architecture, providing an efficient, scalable architecture to facilitate integration of multiple stages of linguistic processing. By June 2001, the IFE-Bio system will provide richer linguistic processing through the integration of modules contributed by other TIDES participants. By June 2002, the IFE-Bio system will include additional functionality, such as real-time broadcast news feeds, new machine translation components, support for question-answering, cross-language information retrieval, multi-document summarization, automatic extraction and normalization of temporal and spatial information, and automated geospatial and temporal displays.

2. The IFE-Bio System

The current demonstration (March 2001) highlights the basic functionality required by an analyst, including:

- **Capture** of sources, including e-mail, digital library material, news groups, and web-based resources;
- **Categorizing** of the sources into multiple orthogonal hierarchies useful to the analyst, e.g., disease, region, news source, language;
- **Processing** of the information through various stages, including "zoning" of the text to select the relevant portions for processing; named entity detection, event detection, extraction of temporal information, summarization, and translation from Spanish, Portuguese, and Chinese into English;
- **Access** to the information through use of any mail and news group reader, which allows the analyst to organize, save, and share the information in a familiar, readily accessible environment;
- **Display** of the information in alternate forms, including color-tagged documents, tables, summaries, graphs, and geospatial, map-based displays.

Figure 1 below shows the overall functionality envisioned for the IFE-Bio system, including capture, categorizing, processing, access and display.

Collection capability for the current IFE-Bio system includes email, news groups, journals, and Web resources. We have a complete copy of the ProMED mailings (a moderated source

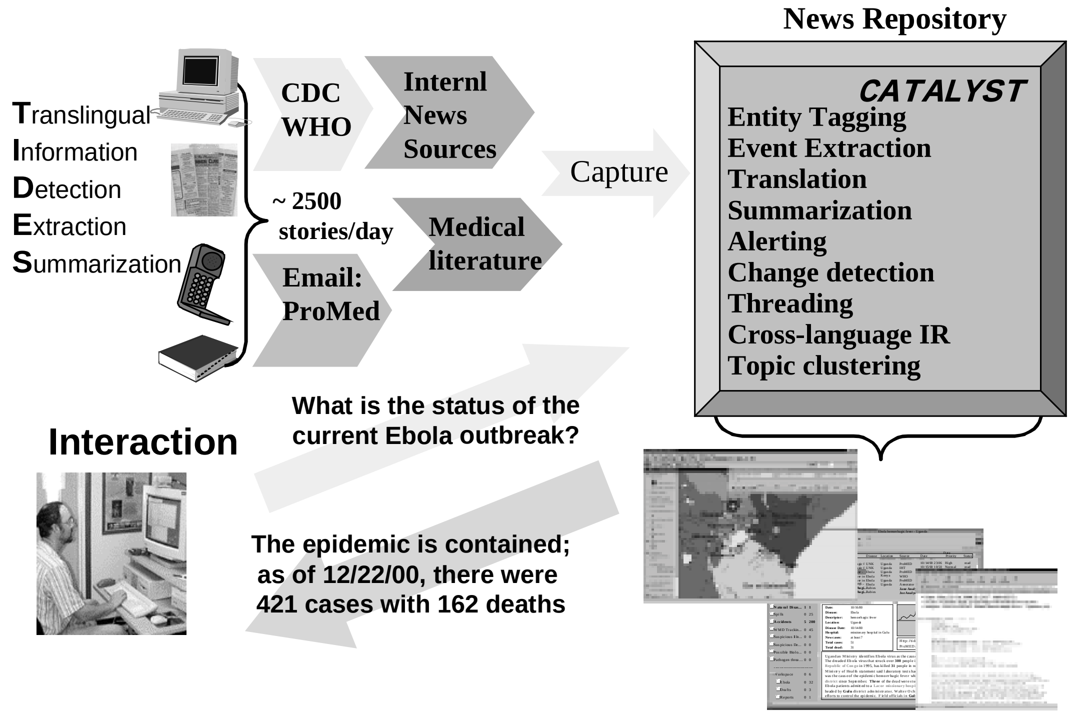


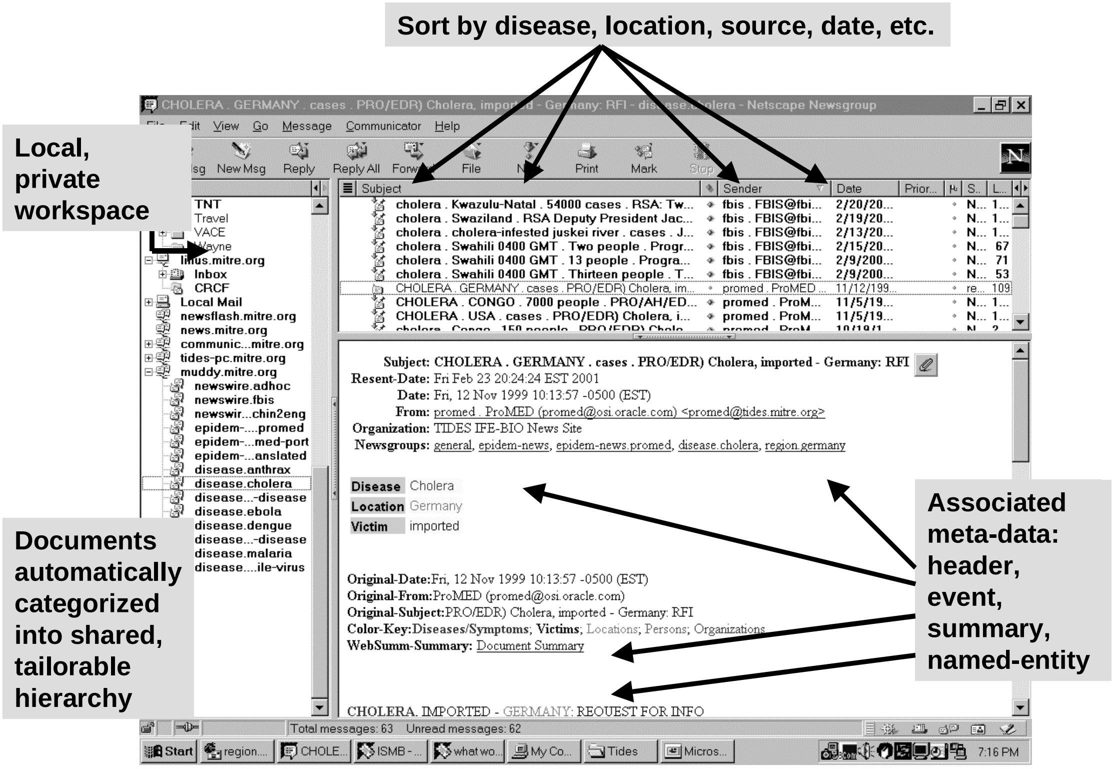
Figure 1: Overview of the IFE-Bio Demonstration System

tracking global infectious disease outbreaks), and are routinely collecting other information sources from the World Health Organization and CDC. In addition, we are collecting several general global news feeds. Current volume is around 2000 messages per day; we estimate capacity for the current system at around 4500 messages/day. Once we have integrated a filtering capability, we expect the volume of messages saved in IFE-Bio should drop significantly, since many of the global news services report on a wide range of events and not all need to be passed on to IFE-Bio analysts. The categorizing of sources is done based on the message header. The header is synthesized by extracting key information about disease name, the country, and other relevant information such as type of victim and source of information, as well as date of message receipt.

The processing for the current demonstration system uses a limited subset of the Catalyst architecture capabilities and a number of in-house linguistic modules. The linguistic modules in the current demonstration system include tokenization, sentence segmentation, part-of-speech tagging, named entity detection, temporal extraction (Mani and Wilson 2000) and source-specific event detection. In addition, we have incorporated the CyberTrans embedded machine translation system which "wraps" available machine translation engines to make them available via an e-mail or Web interface (Reeder 2000). Single document summarization is performed by the MITRE WebSumm system (Mani and Bloedorn 1999).

We carefully chose a light-weight interface mechanism for delivery of the information to the analyst. By treating the incoming streams of data as feeds to a news server, the analyst can inspect and organize the information using a familiar news and e-mail browser. The analyst can subscribe to areas of interest, flag important messages, watch specific threads, and create tailored filters for monitoring outbreaks. The stories are crossed-posted to multiple relevant news groups, based on the information in the header, e.g., a story on Ebola in Africa would be cross posted to the Africa regional newsgroup and to the Ebola disease newsgroup. Search by subject and date allow the analyst to select subsets of the messages for further processing, annotation or sharing. The news client provides notification of incoming messages. In later versions, we plan to integrate topic detection and tracking capabilities, to provide improved filtering and routing of messages, as well as detection of new topics. The use of this simple delivery mechanism provides a familiar environment with almost no learning curve, and it avoids issues of platform and operating system dependence.

Finally, the system makes use of several different devices to display the information appropriately. Figure 2 shows the layout of the Netscape news browser interface. It includes the list of newsgroups that have been subscribed to (on the left), the list of messages from the chosen newsgroup (on top), and a particular message with color-coded named entities (including disease terms displayed in red, so that they are easy to spot in the message).



Subject	Sender	Date	Prior...	S..	L...
cholera . Kwazulu-Natal . 54000 cases . RSA: Tw...	fbis . FBIS@fbi...	2/20/20...		N...	1...
cholera . Swaziland . RSA Deputy President Jac...	fbis . FBIS@fbi...	2/19/20...		N...	1...
cholera . cholera-infested juskei river . cases . J...	fbis . FBIS@fbi...	2/13/20...		N...	1...
cholera . Swahili 0400 GMT . Two people . Progr...	fbis . FBIS@fbi...	2/15/20...		N...	67
cholera . Swahili 0400 GMT . 13 people . Progra...	fbis . FBIS@fbi...	2/9/200...		N...	71
cholera . Swahili 0400 GMT . Thirteen people . T...	fbis . FBIS@fbi...	2/9/200...		N...	53
CHOLERA . GERMANY . cases . PRO/EDR) Cholera, im...	promed . ProMED ...	11/12/199...		re...	109
CHOLERA . CONGO . 7000 people . PRO/AH/ED...	promed . ProM...	11/5/19...		N...	1...
CHOLERA . USA . cases . PRO/EDR) Cholera, i...	promed . ProM...	11/5/19...		N...	1...
cholera . Congo . 150 people . PRO/EDR) Chole...	promed . ProM...	10/19/1...		N...	2...

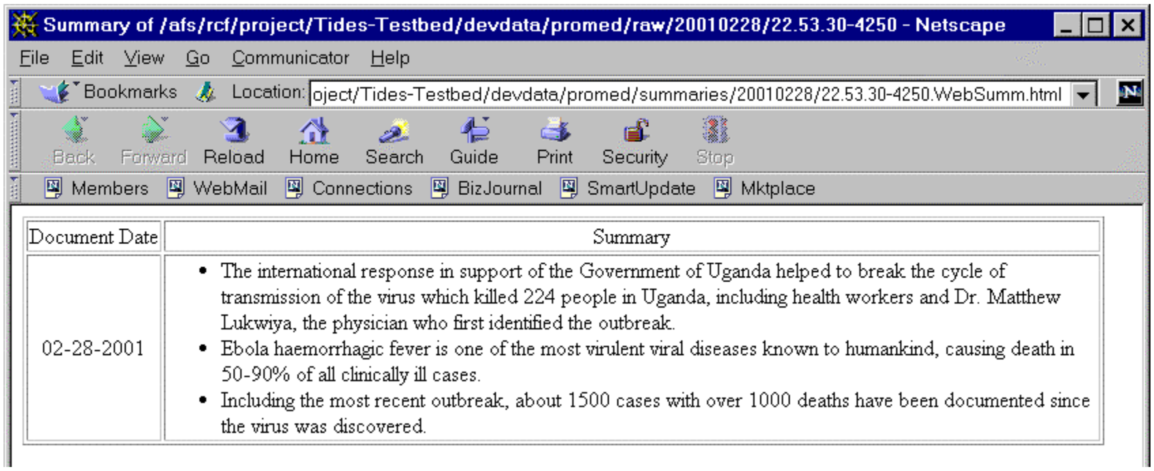
Subject: CHOLERA . GERMANY . cases . PRO/EDR) Cholera, imported - Germany: RFI
Resent-Date: Fri Feb 23 20:24:24 EST 2001
Date: Fri, 12 Nov 1999 10:13:57 -0500 (EST)
From: promed . ProMED (promed@osi.oracle.com) <promed@tides.mitre.org>
Organization: TIDES IFE-BIO News Site
Newsgroups: general, epidem-news, epidem-news.promed, disease.cholera, region.germany

Disease	Cholera
Location	Germany
Victim	imported

Original-Date: Fri, 12 Nov 1999 10:13:57 -0500 (EST)
Original-From: ProMED (promed@osi.oracle.com)
Original-Subject:PRO/EDR) Cholera, imported - Germany: RFI
Color-Key:Diseases/Symptoms; **Victims**; Locations; Persons; Organizations
WebSumm-Summary: Document Summary

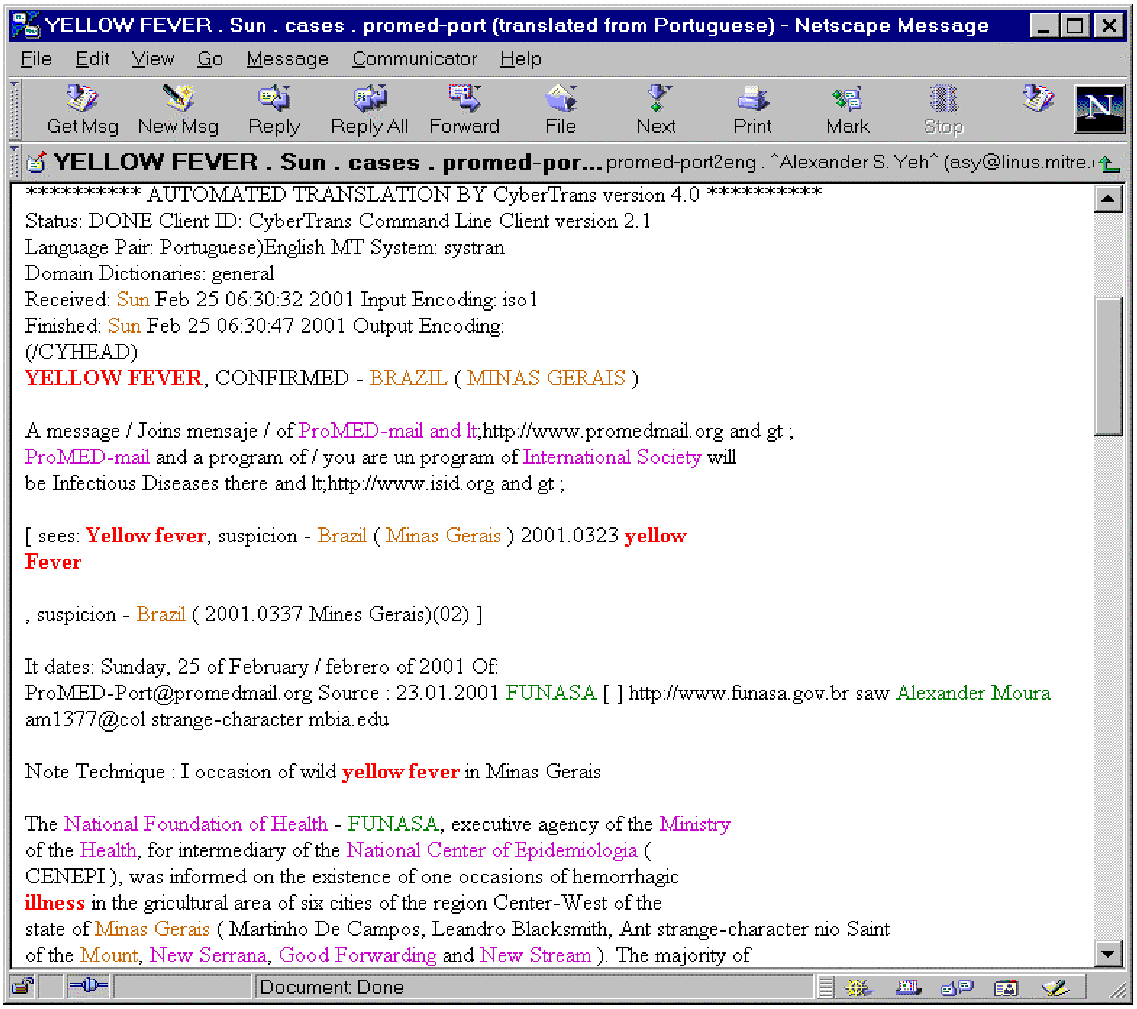
CHOLERA, IMPORTED - GERMANY: REQUEST FOR INFO

Figure 2: Screenshot of IFE-Bio Interface Using News Group Reader


Summary of /afs/rcf/project/Tides-Testbed/devdata/promed/raw/20010228/22.53.30-4250 - Netscape
File
Edit
View
Go
Communicator
Help
Bookmarks
Location:
oject/Tides-Testbed/devdata/promed/summaries/20010228/22.53.30-4250.WebSumm.html
Back
Forward
Reload
Home
Search
Guide
Print
Security
Stop
Members
WebMail
Connections
BizJournal
SmartUpdate
Mktplace

Document Date	Summary
02-28-2001	• The international response in support of the Government of Uganda helped to break the cycle of transmission of the virus which killed 224 people in Uganda, including health workers and Dr. Matthew Lukwiya, the physician who first identified the outbreak. • Ebola haemorrhagic fever is one of the most virulent viral diseases known to humankind, causing death in 50-90% of all clinically ill cases. • Including the most recent outbreak, about 1500 cases with over 1000 deaths have been documented since the virus was discovered.

Figure 3: Sample Summarization Automatically Generated by WebSumm


YELLOW FEVER . Sun . cases . promed-port (translated from Portuguese) - Netscape Message
File Edit View Go Message Communicator Help
Get Msg
New Msg
Reply
Reply All
Forward
File
Next
Print
Mark
Stop
YELLOW FEVER . Sun . cases . promed-por... promed-port2eng . ^Alexander S. Yeh^ (asy@linus.mitre.
Document: Done

*********** AUTOMATED TRANSLATION BY CyberTrans version 4.0 ***********
Status: DONE Client ID: CyberTrans Command Line Client version 2.1
Language Pair: Portuguese)English MT System: systran
Domain Dictionaries: general
Received: Sun Feb 25 06:30:32 2001 Input Encoding: iso1
Finished: Sun Feb 25 06:30:47 2001 Output Encoding:
(/CYHEAD)
YELLOW FEVER, CONFIRMED - BRAZIL (MINAS GERAIS)

A message / Joins mensaje / of ProMED-mail and lt;http://www.promedmail.org and gt ;
ProMED-mail and a program of / you are un program of International Society will
be Infectious Diseases there and lt;http://www.isid.org and gt ;

[sees: Yellow fever, suspicion - Brazil (Minas Gerais) 2001.0323 yellow
Fever

, suspicion - Brazil (2001.0337 Mines Gerais)(02)]

It dates: Sunday, 25 of February / febrero of 2001 Of:
ProMED-Port@promedmail.org Source : 23.01.2001 FUNASA [ ] http://www.funasa.gov.br saw Alexander Moura
am1377@col strange-character mbia.edu

Note Technique : I occasion of wild yellow fever in Minas Gerais

The National Foundation of Health - FUNASA, executive agency of the Ministry
of the Health, for intermediary of the National Center of Epidemiologia (
CENEPI), was informed on the existence of one occasions of hemorrhagic
illness in the gricultural area of six cities of the region Center-West of the
state of Minas Gerais (Martinho De Campos, Leandro Blacksmith, Ant strange-character nio Saint
of the Mount, New Serrana, Good Forwarding and New Stream). The majority of

Figure 4: Translation from Portuguese to English Produced by CyberTrans

There are multiple display modalities available. The message in Figure 2 contains a short tabular display in the beginning, identifying disease, region and victim type. Below that is a URL to a document summary, created by MITRE's WebSumm system (see Figure 3 for a sample summary). If an incoming message is in a language other than English, then CyberTrans is called to run code set and language identification modules, and the language is translated into English for further processing. Figure 4 below shows a sample translated message; note that there are a number of untranslated words, but it is still possible to get the gist of the message.

In addition, we are working on a mechanism to provide geographic and eventually, temporal display of outbreak information. Figure 5 shows the stages of processing involved. Stage 1 shows onamed entity and temporal tagging to identify the items of interest. These are combined into disease events by further linguistic processing; the result is shown in the table in Stage 2. This spreadsheet of events serves as input for a map-based display, shown in Stage 3. The graph plots number of new cases and number of cumulative cases over time. In the map, the size of the outer dot represents total number of cases to date, and the inner dot represents new cases. This allows the analyst to visualize spread of the disease, as well as the stage of the outbreak (spreading or subsiding).

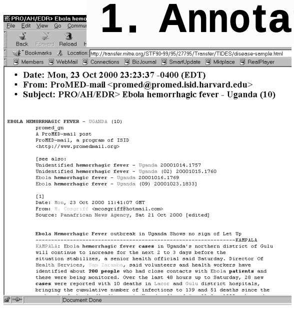
3. REFERENCES

[1] Mani, I. and Bloedorn, E. (1999). "Summarizing Similarities and Among Related Documents". Information Retrieval 1(1): 35-67.

[2] Mani, I. and Wilson, G. (2000). "Robust Temporal Processing of News," Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'2000), 69-76. New Brunswick, New Jersey. Association for Computational Linguistics.

[3] Reeder, F. (2000) "At Your Service: Embedded MT as a Service", NAACL Workshop on Embedded MT, March, 2000.

1. Annotate entities of interest via XML



PRO/AH/EDR> Ebola hemo
File Edit View Go Commu
Back Forward Reload
Bookmarks Location: http://transfer.mitre.org/STF90-99/95/27795/Transfer/TIDES/disease-sample.html
Members WebMail Connections BizJournal SmartUpdate Mktplace RealPlayer
Document: Done

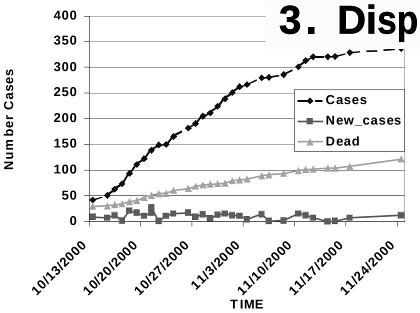
- **Date: Mon, 23 Oct 2000 23:23:37 -0400 (EDT)**
- **From: ProMED-mail <promed@promed.isid.harvard.edu>**
- **Subject: PRO/AH/EDR> Ebola hemorrhagic fever - Uganda (10)**

```
EBOLA HEMORRHAGIC FEVER - UGANDA (10)
          promed_gn
          A ProMED-mail post
          ProMED-mail, a program of ISID
          <http://www.promedmail.org>

          [see also:
          Unidentified hemorrhagic fever - Uganda 20001014.1757
          Unidentified hemorrhagic fever - Uganda (02) 20001015.1760
          Ebola hemorrhagic fever - Uganda 20001016.1769
          Ebola hemorrhagic fever - Uganda (09) 20001023.1833]

          [1]
          Date: Mon, 23 Oct 2000 11:41:07 GMT
          From: M. Cosgriff <mcosgriff@hotmail.com>
          Source: Panafrican News Agency, Sat 21 Oct 2000 [edited]


          Ebola Hemorrhagic Fever outbreak in Uganda Shows no sign of Let Up
          ----------------------------------------------------------------KAMPALA
          KAMPALA: Ebola hemorrhagic fever cases in Uganda's northern district of Gulu
          will continue to increase for the next 2 to 3 days before the
          situation stabilizes, a senior health official said Saturday. Director Of
          Health Services, Sam Zaramba, said volunteers and health workers have
          identified about 200 people who had close contacts with Ebola patients and
          these were being monitored. Over the last 48 hours up to Saturday, 28 new
          cases were reported with 10 deaths in Lacor and Gulu district hospitals,
          bringing the cumulative number of infections to 139 and 51 deaths since the
```

Disease	Source	Country	City_nam	Date	Cases	New_cases	Dead
Ebola	PROMED	Uganda	Gula	26-Oct-2000	182	17	64
Ebola	PROMED	Uganda	Gula	5-Nov-2000	280	14	89
Ebola	PROMED	Uganda	Gulu	13-Oct-2000	42	9	30
Ebola	PROMED	Uganda	Gulu	15-Oct-2000	51	7	31
Ebola	PROMED	Uganda	Gulu	16-Oct-2000	63	12	33
Ebola	PROMED	Uganda	Gulu	17-Oct-2000	73	2	35
Ebola	PROMED	Uganda	Gulu	18-Oct-2000	94	21	39
Ebola	PROMED	Uganda	Gulu	19-Oct-2000	111	17	41

2. Assemble entities into events

3. Display events...



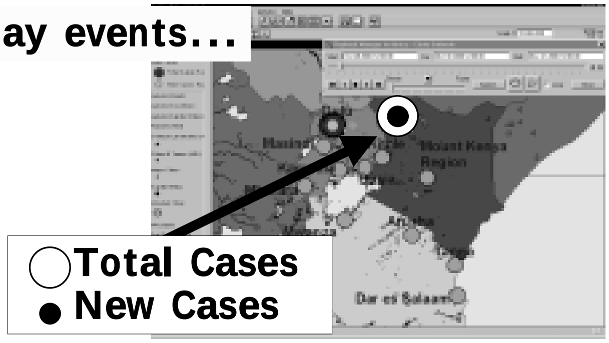


Figure 5: Steps in Extraction to Support Temporal and Geospatial Displays of Disease Outbreak