GALE Data Scouting Task Specification

Version 1.2
Friday, November 11, 2005

Linguistic Data Consortium
http://www.ldc.upenn.edu/Projects/GALE

1 Introduction

The goal of the GALE Data Scouting effort is to find new data types on the Internet that can be harvested for use in GALE. These guidelines describe a process for identifying material that is suitable for use in the program.

In autumn 2004 LDC began very limited exploration of some of these data types – notably, weblogs – by conducting random searches on a known set of websites. The criteria for "good" blogs were that they contain specific mentions of certain kinds of entities, events, and relationships, as defined by the REFLEX Automatic Content Extraction (ACE) project. Random searching proved to be of little value, so these formal guidelines were developed to add structure to the task and to streamline the process.

2 Search Topics

The data scouting effort begins with the Search Topic. Topics are grouped into broad subject headings, drawn from the twelve general event types outlined in the TDT Rules of interpretation. Additional topics were drawn from ACE event types.

For each scouting session, the data scout is assigned a particular subject and topic to concentrate on. Examples include:

> Subject: *Natural Disaster*
> Search Topics: *hurricane, landslide/mudslide, tornado...*
>
> Subject: *Terrorism*
> Search Topics: *ambush, bioterrorism, car bombings, land mines...*

Some Search Topics contain potentially distressing information, which can be emotionally taxing for an annotator to read for hours at a time. If an annotator is upset by a topic or if they believe that mental fatigue is hampering their ability to search, they can note this in the DataScouting Toolkit and ask for a new topic to be assigned.

3 Data Types

Searching is also constrained by asking the annotator to focus on one particular data type per session. Each type is described below. GALE's focus is limited to weblogs and newsgroups.

For each session the annotator is assigned a quota designating the number of "good" sites that need to be found for a given topic and data type in order to consider the session complete.

3.1 WEBLOGS

A weblog (blog) is a shared online journal where people can post diary entries about their personal experiences and hobbies. Some blogs are also used as a forum to express opinions, and the owner of the blog can entertain discussion in the form of posted messages from visitors to the blog. A single blog can include entries from one or many people.

For more information, see the Wikipedia Blog Page.

3.2 NEWSGROUPS/FORUMS/BULLETIN BOARDS

Usenet is a worldwide bulletin board system that can be accessed through the Internet or through many online services. Usenet contains thousands of forums, called newsgroups, which serve every imaginable interest group. Usenet groups can be accessed by anyone, and contain informal messages on a variety of topics, as well as news and information from wire services such as the Associated Press and Reuters News Agency.

A web forum is another kind of internet-based bulletin board system, where people post messages on any number of topics. Forums are the descendants of bulletin board and Usenet systems. Members of the forum create topics, or threads, and other members post their responses under the thread. Forums are sometimes administered by

moderators who alter or delete the posts of members, and limit member access to certain threads.

For more information, see the Wikipedia Usenet Page and the Wikipedia Forum Page.

3.3 WEB NEWS AND MAGAZINES

Web news includes online magazines, newspapers or other data providers whose content is only web-based, for instance, magazines like Slate.com or WorldNetDaily.com. Web-only content from print media organizations like Wired is also considered web news. Web News does not include news content that would also appear in another medium, such as television, radio, or newsprint. Sources like news.google.com that reprint wire feeds from other data providers do not constitute web news. Web news is limited to original content presented solely on the web.

3.4 CHAT ROOMS

Chat rooms are another type of online forum where messages, generally short text messages, are exchanged in real time. Unlike other online forums, chat rooms usually do not archive messages. Some chat rooms are structured and moderated, while others allow anyone to participate. Annotators are not required to go into chat rooms; they simply note the URL where a topicalized chat takes place, and include any other information they can find – particular times the chat is most active, etc.

For more information, see the Wikipedia Chatroom Page.

3.5 LEGAL PROCEEDINGS

Legal Proceedings include courtroom transcripts, indictments, court decisions, testimony and other official documentation issued by the court regarding legal cases. News articles reporting about a trial do not constitute a Legal Proceedings document.

3.6 GOVERNMENT DOCUMENTS

These are documents officially published by a local, state, or federal government. They include laws, statutes, press statements, the text of speeches, debates, briefings, and official reports from government agencies. News reports about such documents are not Government Documents.

4 Content Assessment

Once an annotator has found a document or webpage that fits one of the Data Types and includes information regarding the Search Topic, they judge the document for its content. Good and bad judgments are based on the specificity of the information in the text about the Topic. Good documents generally contain specific detailed events, entities and relations, including name, place, and temporal information that is relevant to the assigned topic.

For example, a chat room where participants debate evidence in the Michael Jackson trial would be judged good for the topic Trials. An editorial ranting about air travel would be judged bad for the topic Vacations, while a blog where people share specific travel experiences would be good for that topic. A forum thread debating whether drug companies should distribute vaccines to Africa for free may contain both good and bad threads depending on how specific or vague the information is. A user posting an opinion like "I think companies should distribute free AIDS drugs in Africa since so many people are dying" is less desirable than a post debating particular policies in place among specific drug companies for distributing AIDS drugs in Botswana. In addition to judging documents as good or bad, annotators may add comments describing the motivation for their judgment.

The DataScout Toolkit also contains an applet that allows the annotator to select individual passages from a larger document and label them as particularly good (or particularly bad)

examples of content for the designated topic. These text strings are logged to the database and can be used to train automatic data scouting and selection processes.

5 Search Process

The annotation process is controlled by an automatic workflow manager, AWS. Annotators log into AWS and are automatically assigned a search topic. AWS also launches the customized DataScout Toolkit and a web browser.

The DataScout Toolkit contains multiple windows to control the data scouting task. The left pane shows a tally of the data types found for the annotator's topic. The top pane is occupied by a web browser, and the bottom pane consists of a window where the annotator inputs information including data type, title, and site URL.

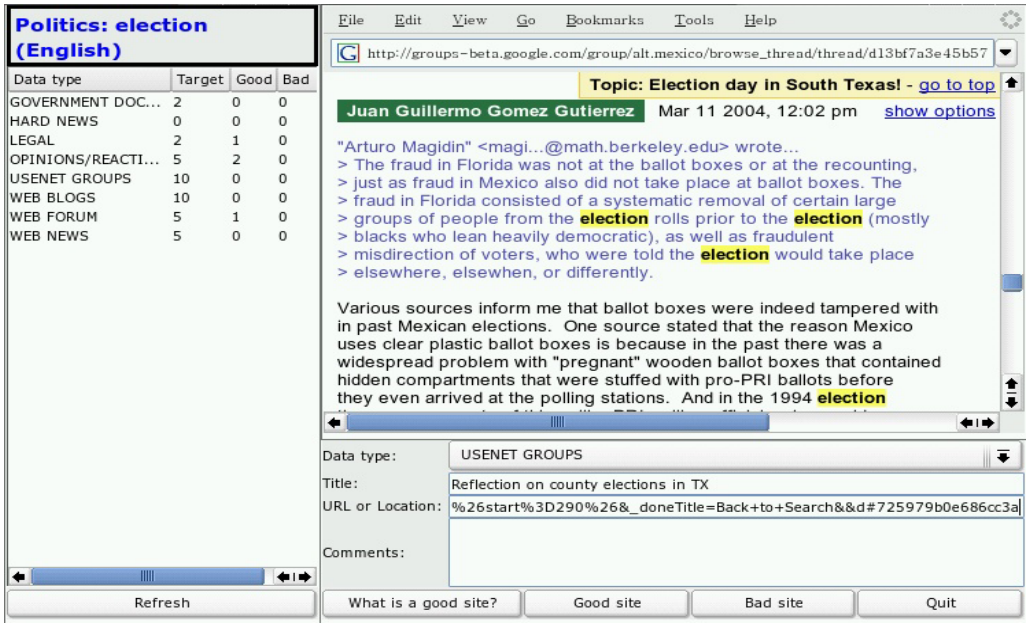

Figure 1. Screen shot of the DataScouting toolkit

The tool also contains help functions that include descriptions and examples of good sites, and a list of recommended search engines that are particularly effective for the targeted data types (see Appendix 2).

6 Quality Control

Websites that have been submitted via the DataScouting toolkit undergo a quality control pass conducted by supervisors, who review submissions and rate the quality of sites at this point. Other quality control measures are conducted during the harvesting and processing stages, where badly formatted documents are weeded out.

7 Data Harvesting and Processing

Once supervisors have approved a site and created a curl pattern for downloading, a nightly process queries the database and harvests all designated URLs.

Data on the web occurs in numerous formats, with highly variable (and inconsistently-applied) markup. We have developed a number of scripts to standardize formatting so data can be more easily fed into downstream annotation processes. Original-format versions of each document are also preserved. Typically a new script is required for each new domain name that has been identified. After scripts are run, an optional manual process corrects any remaining formatting problems.

Once the formatting standards for a particular domain name have been put in place and debugged, we automate harvesting data from that domain name on a daily basis. So for

instance, if scouts identify several "good" blogs all hosted by blogger.com, the automated process would harvest everything in the *.blogger.com domain and apply the blogger.com formatting scripts to that data.

8 Data Selection

Automated harvesting results in a large data pool that contains "bad" as well as "good" documents (as defined by their information content). Before documents are earmarked for downstream annotation tasks they still must be reviewed for content suitability; failure to do so might result in for instance blogs about knitting patterns or vegan recipes being translated or Treebanked.

Data selection is a semi-automated process. Documents and text passages already labeled as "good" provide input to a statistical analysis of token frequency for good/bad documents for each topic or data type. The analysis is used to generate a list of positively- and negatively-weighted keywords to help in the identification of additional "good" documents from the data pool. The list of keywords is then fed through LDC's custom search engine to generate relevancy rankings for each document. Some additional processes exclude easily-identified "junk" documents. Finally, an annotator reviews the list of relevance-ranked documents and selects those which are suitable for a particular annotation task or for annotation in general. These newly-judged documents in turn provide additional input for generation of new ranked lists.

Appendix 1: Examples of Good Sites

An example of a good quality newsgroup is shown below. This thread was judged to be useful for future annotation due to its specific content. The examples show the site both as it appears on the web and as it appears after being harvested and formatted.

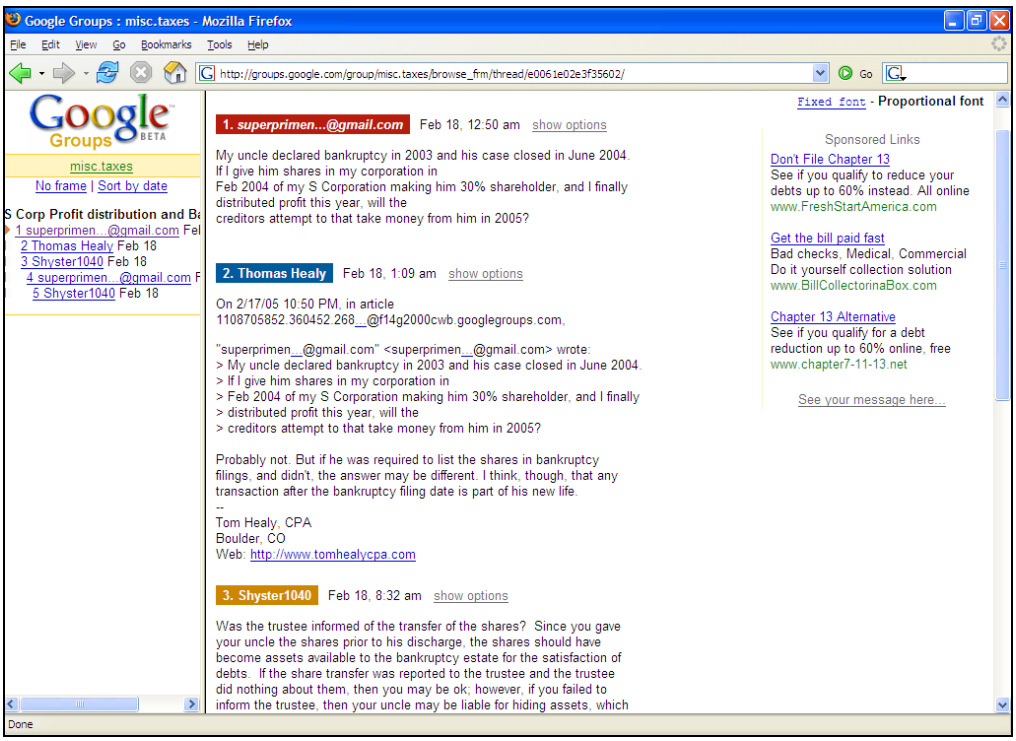

Figure 2. English Newsgroup – Original Format

```
<DOC>
<DOCID>misc.taxes_20050218.1250</DOCID>
<DOCTYPE SOURCE="usenet">USENET TEXT</DOCTYPE>
<DATETIME>2005-02-18T12:50:00</DATETIME>
<BODY>
<HEADLINE>
S Corp Profit distribution and Bankrupcy.
</HEADLINE>
<TEXT>

<POST POSTER="superprimen...@gmail.com" DATETIME="2005-02-18T12:50:00">
Newsgroups: misc.taxes
From: superprimen...@gmail.com
Date: 17 Feb 2005 21:50:52 -0800
Local: Fri ,Feb 18 2005 12:50 am
Subject: S Corp Profit distribution and Bankrupcy.

My uncle declared bankruptcy in 2003 and his case closed in June 2004. If I give
him shares in my corporation in Feb 2004 of my S Corporation making him 30%
shareholder, and I finally distributed profit this year, will the creditors
attempt to that take money from him in 2005?

</POST>

<POST POSTER="Thomas Healy" DATETIME="2005-02-18T01:09:00">
Newsgroups: misc.taxes
From: Thomas Healy &lt;tomhealy...@earthlink.net&gt;
Date: Fri, 18 Feb 2005 06:09:54 GMT
```

```
Local: Fri ,Feb 18 2005 1:09 am
Subject: Re: S Corp Profit distribution and Bankrupcy.

On 2/17/05 10:50 PM, in article
1108705852.360452.268...@f14g2 000cwb.googlegroups.com,

<PREVIOUS_POST "
&quot;superprimen...@gmail.com&quot;&lt;superprimen...@gmail.com&gt; wrote:
&gt; My uncle declared bankruptcy in 2003 and his case closed in June 2004.
&gt; If I give him shares in my corporation in
&gt; Feb 2004 of my S Corporation making him 30% shareholder, and I finally
&gt; distributed profit this year, will the
&gt; creditors attempt to that take money from him in 2005?
">

Probably not. But if he was required to list the shares in bankruptcy filings, and
didn't, the answer may be different. I think, though, that any transaction after
the bankruptcy filing date is part of his new life.
--
Tom Healy, CPA
Boulder, CO
Web:

</POST>

<POST POSTER="Shyster1040" DATETIME="2005-02-18T08:32:00">
Newsgroups: misc.taxes
From: &quot;Shyster1040&quot;&lt;Shyster1...@nospamhotmail.com&gt;
Date: Fri, 18 Feb 2005 08:32:13 -0500
Local: Fri ,Feb 18 2005 8:32 am
Subject: Re: S Corp Profit distribution and Bankrupcy.

Was the trustee informed of the transfer of the shares? Since you gave your uncle
the shares prior to his discharge, the shares should have become assets available
to the bankruptcy estate for the satisfaction of debts. If the share transfer was
reported to the trustee and the trustee did nothing about them, then you may be
ok; however, if you failed to inform the trustee, then your uncle may be liable
for hiding assets, which would, at the minimum, invalidate his discharge. Even if
the transfer doesn't invalidate the whole discharge, the distribution might be
subject to forfeiture to pay creditors' claims if you failed to inform the trustee
of the share transfer.

You need to speak to a bankruptcy attorney pronto; this is a bankruptcy matter,
not a tax matter.

</POST>

</TEXT>
</BODY>
</DOC>
```

Figure 3. English Newsgroup – GALE Standardized Format

```
<DOC>
<DOCID>THELAMEDUCK_20041113.1300.011</DOCID>
<DOCTYPE SOURCE="blog">BLOG TEXT</DOCTYPE>
<DATETIME>2004-11-13T13:00:00-06:00</DATETIME>
<BODY>
<HEADLINE>
They Always Go For The Bad Boys
</HEADLINE>
<TEXT>
<POST POSTER="Brandon" DATETIME="2004-11-13T13:00:00-06:00">
It seems like it happens a lot, most recently with Scott Peterson and “wanna-be” assassin
John Hinckley, Jr. They are America’s most wanted … bachelors. Early before the trial
even began, Peterson received among other pieces of mail, love letters from women across
the country would wanted to woo the alleged killer of his wife and unborn son, and the
man who had a affair on top of it. Actually would it be an affair if she was dead
```

```
already, then Scott would just be a grieving widow on the rebound. This isn't an isolated
incident. John Hinckley, Jr., not to be confused with John Hinckley Sr., is currently in
court trying to get longer unsupervised visits to his parent's home.

"Since his 1982 acquittal, Hinckley has gradually has won permission to leave hospital
grounds, first with escorts and then, for short unsupervised visits with his parents. The
trips have been uneventful." Right now the court is investigating a relationship that
Hinckley had with Leslie DeVeau, a fellow "guest" of St Elizabeth's Hospital. What's even
crazier, than love in a loony bin, was that after her release in 1990, she continued the
relationship for nine more years. Wait a minute, she was released, I thought she wasn't
crazy anymore. To this day Hinckley say's they talk twice a day via the phone and he
still wears the engagement ring that DeVeau gave him. With me being a male, I have never
understood the good girls for bad boy's attraction. To bad Freud isn't here to tell us
what it all means. But maybe I'm just in the loony bin on this one, there is no lust for
bad boys, it's just the women out there pulling for the underdog. For those frustrated
ladies out there who are sick of the personals, may I suggest the police blotter?

</POST>
</TEXT>
</BODY>
</DOC>
```

Figure 4. English Blog – GALE Standardized Format

Appendix 2: Recommended search engines

W E B L O G S (English)

Technorati	Keyword search of over six million blogs, updated in real time	BlogStreet	Search thousands of blogs, ranked and organized by topic	Fagan Finder	Searches through large listing of blog indexing and rating sites
Truth Laid Bear	Ranks 5000 blogs by traffic	Blogdex	Indexes blogs by the news stories they link to	E-Talking Head Blog Directory	Political Blog directory, listed by ideology
Blog and Wiki spotting	Blogs organized by continent (Only Europe has listings, some English, some not)	Eatonweb Portal	A listing of blogs by language and topic	Blogwise	37,000+ blogs listed and searchable

W E B L O G S (Arabic)

Globe of Blogs	Use to search blogs from other countries. Browse by location to find in Iraq, Jordan, Syria, Palestinian Territories, etc.	Arab Blog Count	This blog lists all the blogs (according to this blog) written in Arabic. The list is in no particular order or category. It is particularly useful since it lists not only Arab blogs, but those written in Arabic.	Iraq Blog Count	This blog has all of the known Iraqi blogs listed. Many of these blogs are in *English*; however, it also has useful websites and links to Arabic blogs and websites.
Lebanese Blogger Forum	This blogs lists all known Lebanese blogs.	The Egyptian Blog Ring	Very useful website for Egyptian blogs. Can browse by category (Art & Culture, Entertainment, etc.) or by language (English, Arabic, French).	KuwaitBlogs	Listing of Kuwaiti blogs.
Best Arab Blog Awards	Website devoted to recognizing the best Arab/Arabic blogs. Has listing of many blogs and is useful if searching by categories or region.	Bahrain Blogs	Listing of Bahraini Blogs.	Blog Hub	Search blogs by country.
Planet Arab's Bloggers	Small community of Arabic blogs/bloggers.				

W E B F O R U M S and N E W S G R O U P S (English)

Forum Library	A paid listing of around 500 various forums	LookSmart Message Board Directory	Hundreds of boards sorted by interest	Boardhost: Message Board Listings	A listing of boards hosted by boardhost.com, a popular site for hosting
Google Groups	Google's searchable index of newsgroup postings	LookSmart Newsgroup Directory	Links to over twenty newsgroup directories, sorted by subject	LookSmart Usenet Server Listing	Listing of servers hosting newsgroups

W E B F O R U M S and N E W S G R O U P S (Arabic)

Al Basrah "Selected Sites"	Arabic forums, personal political sites, Palestinian Sites, Islamic sites, and more.	Omduena	Claims to be the largest Arabic network directory	2s2s?	Parent site of http://www.2s2s.com/2s2snet/
Abdul's Small World	"Abdul's" site, containing links to chatrooms, Arabic poetry.	ArabsGate Network	List of sites in Arabic -- chatrooms, forums, etc.?	Faharis.net	Forums
Fares.net	News and forums	66c.com	more sites in Arabic	Adleel.org	Arabic news sites, forums
Alhazmi directory for sites	Directory of numerous Arabic sites.				

C H A T R O O M S (English)

SearchIRC	A searchable, sorted index of over 2,000 IRC chatrooms	Raidersoft Chat Index	Searches 200,000 chatrooms, with option to return only active chats	Google Chat Directory	Index of thousands of chats and chat hosting sites
AnySearch Chat Index	Large index of chatrooms and sites	LookSmart Chat Directory	Large listing of chats and sites, sorted into categories	Chat-o-rama	Hundreds of links to chatrooms, sorted alphabetically

G O V E R N M E N T D O C U M E N T S (English)

FedWorld	Government information on science and technology	FirstGov	Huge collection of links to official government sites	THOMAS Legislative Information	Information and text of legislation in Congress, as well as committee reports
FedStats	All manner of statistics kept by all government agencies	SearchSystems	Searches 23,000 databases of public records, including court filings	DefenseLink	Huge index of Defense related sites
Political Money Line	Summary of Federal Election Commission information	NCSL	Links to all State Legislatures	CIA Factbook	CIA info on demographics
World Health Organization	Statistics and information from the WHO	Google Government	Google's search engine for government agencies	State and Local Government on the Net	Links to state and local governments
National Center for	Health statistics from the Centers				

Health Statistics	for Disease Control

L E G A L P R O C E E D I N G S (English)

FindLaw	Index's text of laws and major court decisions	Hieros Gamos	Listing of state, national, and international law	Washburn School of Law	Listing of laws, legal documents, and some decisions
Legal Information Institute	Premier searchable database of laws and decisions	The Smoking Gun	Ecclectic assembly of official public records on celebrity court cases.		

W E B N E W S (English)

- - - C E N T E R / A P O L I T I C A L - - -

Christian Science Monitor	Current events news and commentary	US News and World Reports	Current events news with some commentary	Newsweek	Newsweek web content hosted by MSNBC
Time	Current events news and commentary	Forbes Magazine	Financial and current events reporting and commentary	EmergencyNet News	Original and borrowed news content on security issues
The Economist	Financial and lifestyle news and commentary	Bloomberg.com	Financial news and analysis	Cnet News	Technology News
Wired	Technology and current events news and commentary	Nature	Science magazine	Scientific American	Science magazine
New Scentist	Science magazine	Variety	Entertainment industry magazine	Entertainment Weekly	Entertainment and gossip magazine
People	Lifestyle magazine				

- - - W E B N E W S - - -

- - - L E F T - - -

Mother Jones	News and political commentary	indymedia.org	Network of independent reporting throughout the world	tompaine.com	Political commentary
Washington Monthly	Political news and blog-style commentary	Unknown News	Collection of underreported stories	Federation of American Scientists	Watchdog of government spending and policy
Common Dreams	Commentary on comtemporary issues	Media Matters for America	Media accuracy watchdog group	The Atlantic	Current events magazine

The New Yorker	Commentary on Politics and Lifestyle	E-Talking Head	Original and borrowed news content	The Nation	Political news magazine
Counterpunch	Political essays	The New Republic	Political magazine	Reason	Political magazine
Slate	Internet political magazine	AlterNet	Alternative news daily		

W E B N E W S (English)

- - - R I G H T - - -

WorldNetDaily	Borrowed news content, along with some original writing	American Conservative	Political commentary magazine	Fox News	Original and borrowed news reporting
Sky News	British news service, with original and wire stories	Hill News	Political news and comment from Washington	Human Events	Current events commentary
National Review	Political commentary magazine	Weekly Standard	Political commentary magazine		

W E B N E W S (English)

- - - GENERAL - -

Santa Fe Library	This is the library's research help page	New York Times Cybernavigator	List of web resorces assebmled for New York Times reporters	RobertNiles.com	Los Angeles Times writer's guide to internet searching resources
Drudge Report	News collection and internet gossip site	LDC Annotator Resource Page	Earlier collection of web resources for speech transcription and annotator topic research		