

VRA Events Data Construction Notes

Released February 25, 2005, with a major revision March 7, 2006

The VRA events data files contain parsed news report output (events). There are 3 files, divided into 5-year sections. The total number of parsed events contained in the three files is 10,252,937 (post-filtering as discussed below). Because we parsed two sentences per report, we have included separate event totals for the first and second sentences.

Table 1. Report and Events Totals

File Name	Totals			
	Reports	Events	Events -- Sentence 1	Events -- Sentence 2
2000-2004	3,159,155	3,464,897	1,977,049	1,487,848
1995-1999	3,770,067	4,108,102	2,412,053	1,696,049
1990-1994	1,914,095	2,679,938	1,627,147	1,052,791

News Source

The news report source for all of the files is Reuters; however the specific product lines differ slightly, depending on month and year. From 1990 to May 2003, we parse the Reuters Business Briefings (RBB). From June 22, 2003 through September 9, 2003, the report source parsed was Factiva's World News. Since September 10, 2003 the report source is Reuters (World) news, as carried in eight of their product channels.

Processing Details

These files were parsed using version 3.11.0 of the VRA Reader program. We set the parser to output two sentences per report (the first and the second sentence). Processing began on 1/27/2005 and ended on 2/7/2005.

Filtering Details

The filters we employ address three basic issues as follows:

1. Format - We filter out reports that are *not* news reports; the most common (see the table below for a full count) of these include tables, digests and research alerts. We employ this filter because the parser is optimized to process news reports.
2. Content - We filter out routine stock market (typically opening and closing) reports, exchange rates and shipping schedules. We employ this filter because they are regular (almost a constant) feature of reporting in industrialized countries that is not indicative of trends except in their content. Please note, however, that we are experimenting with these reports in an attempt to process the content in a way that captures the economic trends they track. We may very well be able to include them in the next public release of our data. Other content that we exclude include performance (arts and entertainment) and sports events reports/reviews.
3. "Duplicates" - We have made a first attempt in this (1990-2004) data set to minimize duplicates in the sense that the report is "about" the same topic. Note that these entries are not duplicates in the data record sense (all of them are unique records), but the number of typographical corrections and

minor (revising a source's title, for example) updates that follow a report's first posting often runs to a dozen or more.

We envision getting to the point where users can eventually choose to run the data with three options: **first**, the data is presented "as is," with the implication that multiple reports about the same topic are included (but again, not in the data record sense); **second** the original report is presented but all updates and corrections excluded, with the implication of yielding a single report about each topic;" and **third**, only the final corrected version or update on all reports is presented, with the main implication being that the report's date-time stamp will reflect the date/time of the most recent correction or update rather than the date/time of its original release.

For this (1990-2004) release of the IDEA data set, we chose the middle route, so that users are receiving records that are unique in the strict sense, as well as conceptually distinct in their focus. We are considering implementation of the third option for the next public release of the data.

The filtered report (excluded) counts for the 1990-2004 IDEA data set follows:

Filter	CountOfID
table	248619
digest	129228
update	89348
stocks	88553
research alert	79881
parent forecast	63204
corrected	58180
diary	53805
indices	50830
lme	50307
parent results	48080
dollar rates	45721
indicator	40471
euro rates	39668
group forecast	37745
share indices	36912
world cross	36517
rates	
market dollar rate	35761
group result	30511
soccer-	27328
interview	25891
highlights	17893
new issue	16058
world news	14376
weather	14040
schedule	12389
feature-	11469
volume/open	11118

government list	10488
estimated volume	10181
market rate	8415
factors to watch	7590
news summary	6577
text-	5865
hot stocks	5848
cricket-	5757
closing price	5691
tennis-	5493
olympics-	5234
grain/oilseed	5056
money market rates	4848
call money	4682
index future	4379
factbox	4104
instant view	4067
snapshot	4037
opinion	3712
key economic indicator	3709
port condition	3700
initial public offering	3660
=2	3360
full text	3028
quote of the day	2935
advisory	2890
golf-	2692
Indonesia	2552
shipping historical calendar	2323
***	2230
reuters sports	2002
week ahead	1954
baseball-	1699
general and political events	1689
racing-	1686
warrant rate	1624
newsmaker	1578
chronology	1565
vegetable oil prices	1551
Pakistan shipping	1423

column	1409
bond trade	1397
sterling rates	1382
sugar prices	1380
overnight rate	1379
=3	1372
reference rate	1342
eurobond rates	1328
motor racing-	1305
copper	1303
afternoon	
stocks to watch	1253
certificated	1249
gold fixing	1243
physical palm oil	1231
issue log	1224
gold afternoon	1215
gold midday	1214
NBA-	1193
rugby union-	1177
reference rates	1173
shareholding notice	1134
oilseeds price	1122
sports	1112
schedule	
oil product trade	1087
list of	1080
meal prices	1029
cycling-	1024
nhl-	1016
margin index	1014
canola board margin index	1012
athletics-	996
exchange noon	978
european feeds	959
key stock and currency	956
barge freights	916
copper fix	898
in brief	887
jute report	857
business- news-schedule	850
skiing-	840
India shipping	831

wce	786
coffee indicator	774
weekahead	767
features	747
schedule	
Rugby-	737
soybean stocks	731
news graphics	681
repeat-	662
dollar forward	607
swimming-	559
auction price	534
nfl-	527
soy futures	491
prices	
horoscope	471
rallying-	392
quiz	369
sport-	369
agrokhleb	352
away on	347
business	
good morning	334
sailing-	328
boxing-	321
economy-	314
hockey-	299
motorcycling-	298
guest column	273
preview-	273
news in brief	266
sports features	265
schedule	
basketball-	258
eurobond new	251
issue index	
livewire	245
reference	243
(prime) rates	
earnings	233
schedule	
forthcoming	214
world elections	
games-	203
following are	202
odds and ends	191
cobalt web	182
price	
pluggedin	167
shipping	164

update	
badminton-	158
rugby league-	153
doping-	148
shipping	146
condition	
Romania	125
shipping	
refile	118
deliverable	111
wheat and corn	
grain	97
transportation	
Korea shipping	96
investor profile	92
skating-	87
programs:	76
stocksvew	61
reuters data	60
newbiz	59
snooker-	58
special report	54
reuters	50
olympics	
rpt-schedule-	48
Taiwan	46
shipping	
lifting the lid	39
corporate	27
actions	
jumping-	27
statbox-	16
earning result	2
Arab shipping	2
analyst_s view	1

Changes in Reader version 3.11.0

Version 3.11.0 of the VRA Reader includes several modifications, both at the event typology level and at the protocol level. These changes and enhancements are intended to improve not only the range of possible analysis, but also the reliability of the coded data.

Event Typology-Level Changes

At the event typology level, we have made refinements to and collapsed several event categories. Several former events, therefore, have been replaced altogether (such as administrative sanction and denounce categories), while other event forms have been adapted to more specific categories (for example, the creation of suicide and vehicle bombing categories). Table 2 presents event forms that no longer exist as discrete types, but rather have been collapsed or refined into other event types.

Table 2. Event Typology Changes -- Collapsed/Refined Event Forms

Formerly	New Category/ies	Notes
Administrative Sanction	Sanction	Sanctions and administrative sanctions are indistinguishable in news leads.
Biological Weapons Use	Chemical/Biological Weapons Use	The distinction between chemical and biological weapons is often unclear in news leads. Therefore, we collapsed this category to include both.
Bombing	Among Artillery Attack, Small Arms Attack, Suicide Bombing, Mine Explosion, Vehicle Bombing, or Missile Attack categories	This category was expanded to make further refinements to the full range of bombings that are reported.
Commercial Adjustment	Executive Adjustment	Commercial adjustments were collapsed into the Executive Adjustment category.
Denounce	Criticize or Blame	Denunciations were collapsed into the Criticize or Blame category as there was too little differentiation between these categories.
Informational Protest	Protest Demonstration	Informational protests were collapsed into the Protest Demonstration category as this differentiation was rarely distinguishable in news reports.
Military Engagement	Armed Battle	This category was collapsed into Armed Battle as the differentiation was rarely distinguishable in news reports.
Military Seizure	Among Seize Possession, Arrest and Detention, or Abduction categories	In practice, this category was indistinguishable from existing seizure (21) categories.
Other Seizure	Among Seize Possession, Arrest and Detention, or Abduction categories	In practice, this category was indistinguishable from existing seizure (21) categories.
Police Seizure	Among Seize Possession, Arrest and Detention, or Abduction categories	In practice, this category was indistinguishable from existing seizure (21) categories.

Table 3 presents categories that are newly created and constitute wholly new categories. Several of the negotiation/mediation-specific event categories have been created to maintain parity with the Conflict and Mediation Event Observations (CAMEO) data typology schemes.

Table 3. Event Typology Changes -- New Event Forms

New Event Form	Category	Notes
Missile Attack	Force Use	Refined event category to capture more detailed event information.
Agree to Mediation, Agree to Negotiation, Agree to Peacekeeping, and Agree to Settlement	Agree	New event categories added to maintain consistency with CAMEO event typology.
Chemical/biological Weapons Use	Force Use	Refined event category to capture more detailed event information.
Extreme Climactic Condition, Drought, Earthquake, Flood, Hurricane, Tornado, Volcano, Tsunami, Wildfire	Natural Disaster	New event categories added to capture more detailed natural disaster information as contained in news reports.
Demand Information, Demand Policy Support, Demand Aid, Demand Protection, Demand Mediation, Demand Withdrawal, Demand Ceasefire, Demand Meeting, Demand Rights, Investigate Human Rights Abuses, Investigate War Crimes	Demand	New event categories added to maintain consistency with CAMEO event typology.
Corporate Earnings, Corporate Earnings Up, Corporate Earnings Down, Equity Up, Equity Down, Interest Rates, Interest Rates Up, Interest Rates Down	Economic Status	New event categories added to capture more detailed economic report content.
Halt negotiation, Halt Mediation, Reduce or Stop Economic Assistance, Reduce or Stop Humanitarian Assistance, Reduce or Stop Military Assistance, Reduce or Stop Peacekeeping	Sanction	New event categories added to maintain consistency with CAMEO event typology.
Mediate Talks, Engage in Negotiation	Discussion	New event categories added to maintain consistency with CAMEO event typology.
Offer to Negotiate, Offer to Mediate	Propose	New event categories added to maintain consistency with CAMEO event typology.
Ratify a Decision	Endorse	
Reject Ceasefire, Reject Peacekeeping, Reject Settlement, Reject Request for Material Aid, Reject Proposal to Meet, Reject Mediation	Reject	New event categories added to maintain consistency with CAMEO event typology.
Request Mediation, Request Investigation, Request Ceasefire or Withdrawal	Request	New event categories added to maintain consistency with CAMEO event typology.
Release or Return Person, Release or Return Property	Release	Refined event categories to capture more detailed information.
Threaten Boycott or Embargo, Threaten Biological or Chemical Weapons Use, Threaten to Halt Negotiation, Threaten to Halt Mediation, Threaten to Break or Reduce Relations, Threaten to Reduce or Stop Aid	Threaten	New event categories added to maintain consistency with CAMEO event typology.
Security Alert	Warn	New event category to capture more detailed information.

Protocol Level Changes

Systematic changes have been made to the protocol to accommodate the new and enhanced event forms, including the addition of nearly ten thousand new lines of protocol (the dictionary now includes over 25k entries). Notably, the addition of the new economic-oriented event forms necessitated considerable protocol development. We have also worked to improve existing event form reliability by further refining force, protest, and seizure event categories. Much effort was also taken to address false positives in these categories.

Several new entries were also added to the sense index to improve source and target identification. The bulk of the new sense entries are within the business and economic phenomena categories. With the creation and refinement of economic and other categories, it became abundantly clear that modifications to the sense index were necessary. To complement our event and protocol development efforts, we also expanded the number of noun classes to match many of our event categories (for instance, we have added noun classes for suicide bombings and suicide bombers so that we may rely upon fewer literal constructions to code to the recently added suicide bombing category).