GLOBAL ENTREPRENEURSHIP EVENT DATABASE PROJECT

AEL GRANT FINAL REPORT

KALEV LEETARU

OVERVIEW

This report presents the findings of a small pilot study funded by the Academy for Entrepreneurial Leadership (AEL) at the University of Illinois in 2009 to test the feasibility of constructing a global historical database of entrepreneurship events from general newspaper reports. The results of the pilot study are promising and suggest there is real potential in this approach. At the same time, further work will be required to refine the data sources and protocol used for the project. Conversations with College of Business faculty suggest that once these issues are resolved, the database that would be produced through this project will have a revolutionary impact on the study of global entrepreneurship.

DATA SOURCES

The underlying Social Political Economic Event Database (SPEED) project upon which this pilot is based uses the New York Times and Wall Street Journal newspapers as its primary data source. While it might be expected that the Wall Street Journal would provide substantially more events than the New York Times, the Journal ended up providing only slightly more: 60% of all events with the remaining 40% from the Times.

In addition to these two US papers, the Cline Center has almost completed digitization of the complete 1946-2004 runs of the Foreign Broadcast Information Service (FBIS) and Summary of World Broadcasts (SWB). Both are government-operated "open source intelligence" products that monitor radio, television, newspaper, trade journal, and all other open and "grey" literature in each country. Monitored content in the areas of political or economic topics are selected by analysts to offer a representative sample of the media in that country and are transcribed and translated into English. Care is given to ensure that translated content matches the tone and nuances of the original material. It is expected that these collections will yield substantially better coverage of economic activities with foreign nations, though further work will be required to determine the degree to which small-bore entrepreneurial activities are captured compared with large-bore large corporation events.

A project directed by College of Business Professor Rajshree Agarwall on innovation in the medical devices market suggests that the use of press release and business news wires, combined with industry trade publications, yields the richest source of information on small-bore entrepreneurial events. In particular, industry trade publications encapsulate significant trade-specific knowledge on corporate activity and can act as an "index" into those companies that underwent specific types of actions, in order to direct media analysis to collect details on the process of those actions. Even smaller companies often issue press releases, and business news wires try to provide information on a wide range of corporate activities for their clients. Taken together, this information portfolio offers a much richer small-bore view of economic activity that is likely more in line with the needs of entrepreneurship event coding compared with large national newspapers like the New York Times or Wall Street Journal. However, even these sources tend to be heavily focused on economic activities of interest to Western clients, and so it is likely they will need to be supplemented with sources such as FBIS and SWB to offer a truly international perspective.

THE PROCESS

As described in the proposal, a random sample of news articles from the New York Times and Wall Street Journal were selected and coded using the computer model and protocol developed under the previous Cline Center AEL grant to PI Peter Nardulli. Trained human coders were used to read each news article and apply the protocol to code its information into 263 different fields (not all fields are applicable to all events). A total of 640 events were coded in all.

GEOGRAPHIC FOCUS

It is clear that the New York Times' and Wall Street Journal's coverage is strongly biased towards economic activity in the United States: 87% of the events collected were based in this country. This is surprising, as earlier work at the Cline Center on capturing societal stability events suggested that international coverage would be far better represented. Of the foreign events coded, nearly half were at the national-level only, compared with just 18% of US events. In the United States, 25% of events were at the state-level, while 57% were at the city resolution. The 318 US city-level events fell into 127 distinct cities, with the top ten cities listed below. New York City tops the list, which fits its role as the financial capital of the country, but the low number of events from Silicon Valley (11 from San Francisco and zero from San Jose) is concerning, given the role the Valley played in the Dot Com Boom of the 1990's. In particular, the Times' and Journal's close physical proximity to New York City may partially explain this bias. Silicon Valley's storied role in the history of late twentieth century entrepreneurship, combined with its underrepresentation in this database, presents some concerns regarding the validity of using Times or Journal coverage to measure US entrepreneurship.

City	Number Events
New York City	73
Chicago	23
San Francisco	11
Houston	8
Cleveland	7
Philadelphia	7
Dallas	7
Stamford	7
Washington	7
Boston	7

Table 1 - Top 10 US Cities

London dominated the foreign city list, with 30% of all foreign city-level matches. London is indeed one of the financial capitals of Europe, but tends to be more Western-aligned and its significant representation in foreign city-level events suggests a strong Western focus even in foreign event coverage. Only four US states received more than 10 state-level events each, with California having the most (20), followed by Texas (17), New York (14) and Florida (10).

One of the premises of this project was to measure the degree to which national-scope US newspapers like the New York Times and Wall Street Journal capture small-bore events in foreign countries. It is clear that foreign entrepreneurship activity is dramatically underreported in these papers and that most coverage occurs only at the gross national level. Event data derived from these two papers is therefore primarily useful only in measuring activity in the United States, with a strong bias towards East Coast entrepreneurship, even when compared against the innovation centers of the West.

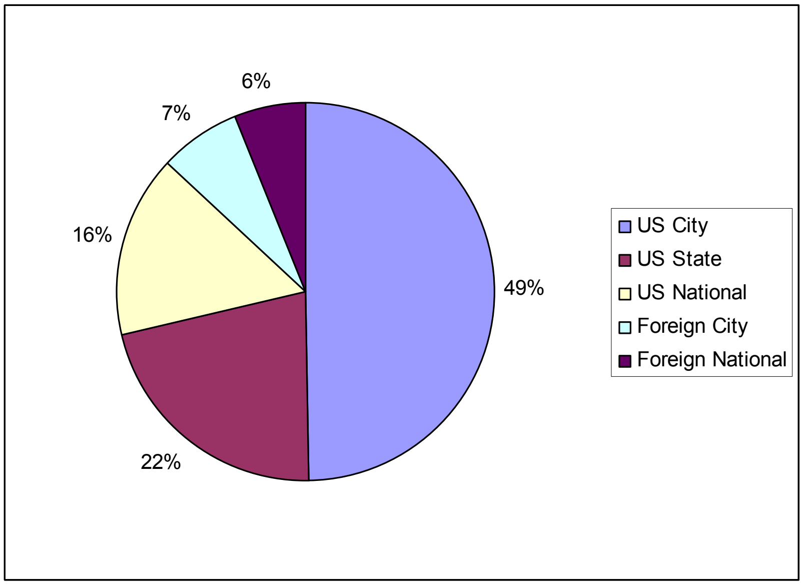

Figure 2 - Percentage of events at each geographic resolution

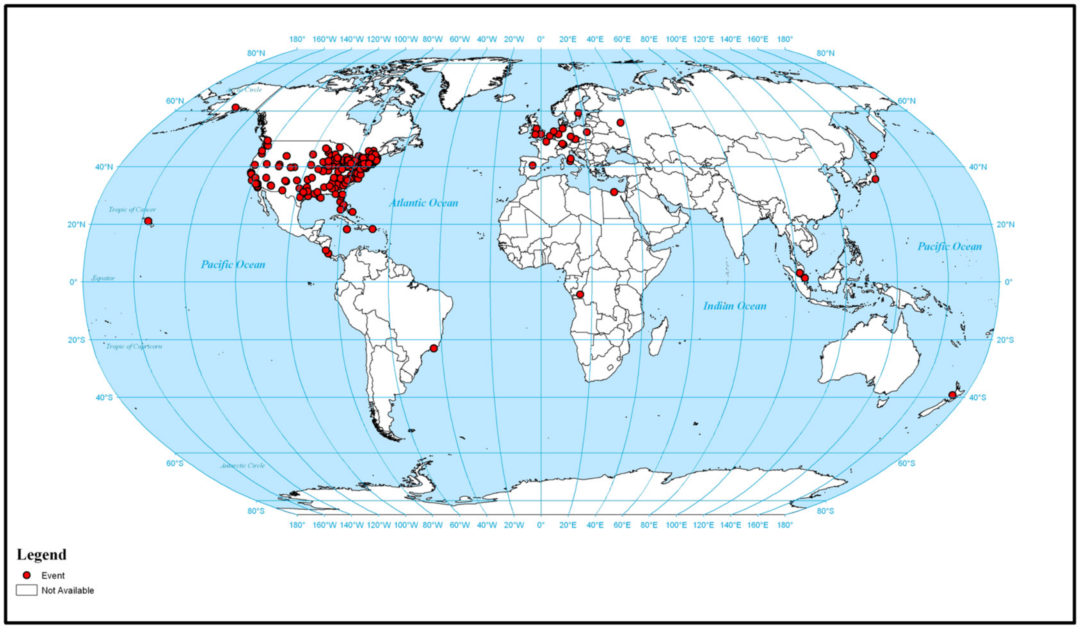

Figure 3 – Event Locations: Global View

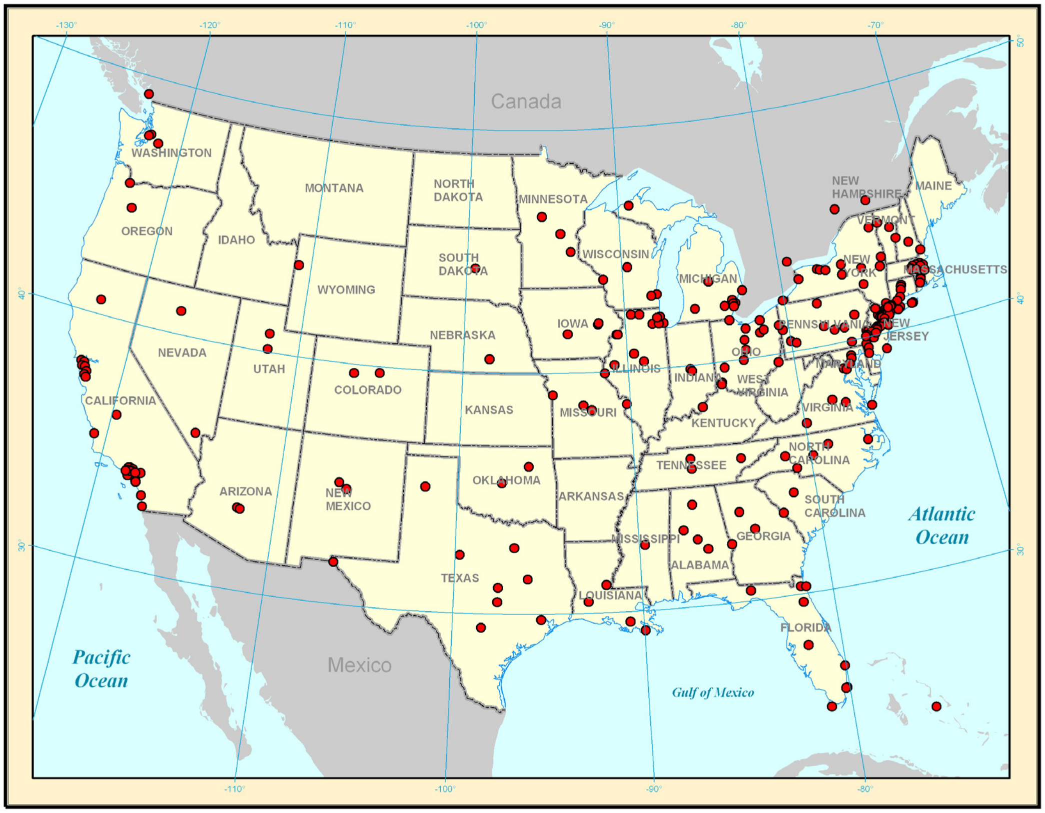

Figure 4 – Event Locations: US View

TEMPORAL FOCUS

Examining event data by decade, a nearly linear increase in event volume is visible from the 1940's through the 1980's. The 1990's, however, represent a 71% decline in entrepreneurship events. Data for the 1940's and 2000's are incomplete, since the underlying news data begins in 1946 and terminates in 2005. Not all articles flagged by the computer contained relevant entrepreneurship events, as noted earlier, so the number of articles per decade is higher than the number of actual coded events.

The substantial decline in the 1990's is unexplainable from an entrepreneurship perspective, in that it was during this decade that the so-called Dot Com Boom occurred, and therefore should contain the highest number of entrepreneurship events of any of the seven decades. One possible explanation lies in the fact that the underlying Wall Street Journal data feed changes to a different provider beginning with content from 1990. News material for the period 1946-1989 comes from the Proquest Corporation and was digitized from historical microfilm copies of the Journal. Beginning in 1990, the Journal launched electronic distribution of its content to commercial aggregator services and Proquest's holdings rely on this data feed for the period 1990-2005. While the "born-digital" Wall Street Journal news feed should be indistinguishable from the feed that was digitized from the print paper, it is possible that there is some form of undetected systematic difference between the data sources.

As noted earlier, the Wall Street Journal is the source for 60% of the events coded in this pilot. Even if the Wall Street Journal content for the period 1990-2005 was completely absent, that should represent a drop of at most 60%, while the actual decline was 10% higher. One possible explanation may be that the predominate language of entrepreneurship-related articles in the 1990's shifts from previous decades. Many of the articles of the 1990's and 2000's focus on technology and internet-driven companies and make use of wording that was not widely used in earlier decades. The underlying computer models used by the categorization system were trained during the previous AEL grant to PI Nardulli using a relatively small number of sample articles, and this may have caused them to become overly sensitive to the language of industries that dominated the landscape in earlier decades. Computer models are very sensitive to language differences, as they do not have the background knowledge of human analysts to fall back upon.

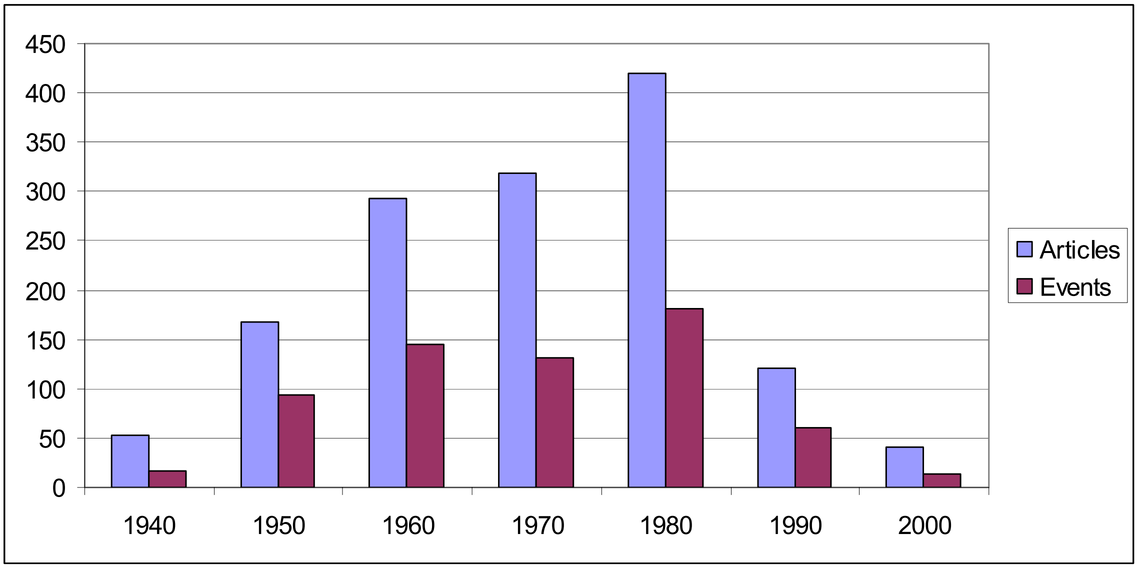

Figure 5 - Number articles/events per decade

DATA ANALYSIS

A total of 620 of the 640 events (96.88%) were classified as "entrepreneurship," while four were "other" and the remaining 16 were "creative destruction." Of entrepreneurship events, 387 (60%) were changes involving more than one entity and 143 (22%) were self-contained changes. None of the remaining event types accounted for more than 3%, and social entrepreneurship and creation of new markets each received less than 1% of cases. Twelve of the 16 creative destruction events (75%) were bankruptcies, while the remaining four were layoffs. It is clear that the current combination of the NYT and WSJ newspapers and the previously-trained selection system will not be capable of capturing a substantial portion of the global bankruptcy activity.

Roughly 60% of events contained information on the affected entity's scope of operations, and 65% of those were domestic (the other 35% were multinational). Slightly more than 99% of events contained ownership information, of which 93% were private corporations, and 3% were public companies. More than 78% of events contained information on the entity's industry of operations, with the largest portion (15%) being manufacturing, 13% finance/insurance, 11% mining/quarrying/oil/gas, and 8% transportation/warehousing. Three quarters of events contained information on the timeframe of the action, with a quarter being proposed future, a quarter imminent future, and a quarter previously occurring. Information on ownership, industry, and timeframe of action are therefore well-represented in the news media.

Cost/value information on deals are available in only a few percent of cases, while funding source is provided for 40%, with stock issue being the most common. Only 1% of records involved entity relocations. Roughly 35% of events contain setting information, with new product/market creation being the most common at 11%, followed by profit decline at 6%. Information on the objective of the action is available for around half of all events, with market expansion the most common at 32%, though the majority of objective information (70%) was inferred by the coder, rather than stated by the companies involved or provided by the news report. Hostility status is included for 40% of events, but only in six cases were the actions deemed hostile. Monopoly and bailout cases appeared in only a few percent of the event data.

Multientity actions involve multiple companies, often at multiple geographic locations, and the protocol includes fields to capture the locations of up to three companies at the city-level resolution. Nearly half of all events included first and second company location information. City-level locative information enables a rich portfolio of spatial analytical techniques to explore geographic entrepreneurship patterns such as diffusion, and this information is well-represented in the news media. Supplementing news reports with additional incorporation and/or registration data sources, such as Hoovers or OneSource, would allow location information to be determined for an even larger set of companies.

ENTREPRENEURSHIP EVENT TYPES

A random sample of 50 multi-entity events suggests that the vast majority of these fall under the heading of mergers and acquisitions. A further sample of self-contained events suggests that a certain percentage of these also include M&A activity as part of a larger growth or diversification strategy. Potentially upwards of three quarters of all events coded under this protocol therefore represent mergers of two or more companies or acquisitions of another company or one of its brands. M&A activity, however, is not considered by most scholars to be "entrepreneurship." While it can create new markets or lead to better efficiency, M&A actions, especially those of large firms, are classified as traditional business ventures, rather than entrepreneurship. The protocol should be revised to more clearly delineate activities involving mergers or acquisitions, or such activities should be excluded from the protocol entirely to place it more in line with existing scholarship.

Social entrepreneurship appears to have been poorly defined in the original protocol, leading to considerable confusion among the human coders. While the number of social entrepreneurship events coded was small (only 8 identified), they do not appear to fit the traditional definition of such activity. For example, articles involving union activity on behalf of wages or benefits are classified as social entrepreneurship, including one 1962 article about a group of six teenage girls in Roselle, New Jersey establishing an informal labor union demanding "work rules and minimum pay standards" for the baby-sitting industry. The American National Theatre and Academy, a federally-chartered organization for the support of the arts, purchased its first headquarters building in 1950, and this too is coded as social entrepreneurship. National social programs in the United States and several other countries such as welfare or home purchasing assistance, are also captured as social entrepreneurship. While the social focus of each of these events is clear, none of them would seem to fall under the traditional interpretation of "social entrepreneurship" and further development and coder guidance for this category is required.

CONCLUSIONS AND FUTURE WORK

It is clear that the event database that would be produced through this project has substantial potential to revolutionize the study of entrepreneurship, but that further work is needed to refine both the protocol and the underlying data sources. Refinement of the protocol instrument is required to address concerns regarding social entrepreneurship and merger and acquisitions behavior, as well as to evaluate the structure of low-performing collection fields. The New York Times and Wall Street Journal appear unable to adequately serve as a source for small-bore entrepreneurship events even in the United States and have severely restricted foreign coverage. New data sources must be located to supplement these two papers. Yet, other projects on campus such as College of Business Professor Rajshree Agarwall's innovation study in the medical devices industry suggest that the approach is sound and that data sources exist which provide the requisite level of small-event coverage.

A follow-on study is therefore recommended that will explore the following four key issues raised in this pilot study:

- **Revision of Core Event Types.** The original protocol developed under the first AEL grant to PI Nardulli included merger and acquisition activity in its definition of entrepreneurship activity. The view of entrepreneurship held by the majority of the business and economics community, however, argues that traditional large-firm M&A activity is not a component of entrepreneurial activity. Up to 75% of the events coded under this pilot treat M&A as entrepreneurship, though it is unclear whether this is an artifact of the newspapers used or a selection bias issue due to the definitions established in the original protocol. A stronger definition of "self-contained" entrepreneurship will be required, as the current definition allows for M&A activity to be coded under self-contained events if they are part of a larger growth or diversification strategy. Given the extremely high proportion of events falling under the M&A heading, it seems useful to modify the protocol to explicitly separate these from other types of entrepreneurship activity, or remove them entirely to bring the protocol more in line with accepted norms in entrepreneurship scholarship.
- **Clarification of Social Entrepreneurship.** While social entrepreneurship was poorly represented in the data captured, it is clear that a more rigorous definition of the topic is needed to improve coder accuracy. Government reforms and official presidential actions were all coded as "social entrepreneurship" activities.
- **Explore New Data Sources**. It is clear that the combination of the Wall Street Journal and New York Times offers a substantially US-biased view of entrepreneurship. The FBIS and SWB intelligence products offer much stronger international coverage, but further work will be required to determine how well they capture small-bore entrepreneurial events. Press release and news wires, together with industry publications, provide another potential source portfolio.
- **Further Training of Computer Categorization Models**. The original computer models used to pre-filter articles for human coding were trained during the first AEL grant using only a few thousand sample articles. In contrast, the Cline Center's production computer models that are used to identify societal stability events were trained using a pool of nearly 100,000 articles. These models are able to achieve 98% accuracy without the decade-level biases seen here, so it is believed that with further training, the accuracy of the selection models could be substantially improved and the 1990's selection artifact can be eliminated. Given that models can often be sensitive to the underlying data feed, work on building more refined models would need to occur after the suitability of new data sources has been evaluated.