Triangle Genesis Data System: Enabling Research on Regional Economics and Innovation

A RENCI Technical Report

TR-17-01

Oleg Kapeljushink, Sidharth Thakur, Karamarie Fecho & Charles Schmitt
Renaissance Computing Institute (RENCI)
University of North Carolina at Chapel Hill

Maryann P. Feldman
Department of Public Policy
University of North Carolina at Chapel Hill

Nichola J. Lowe
Department of City & Regional Planning
University of North Carolina at Chapel Hill

Corresponding author: Maryann P. Feldman, UNC Public Policy, CB#3435, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, maryann.feldman@unc.edu; 919.962.0674

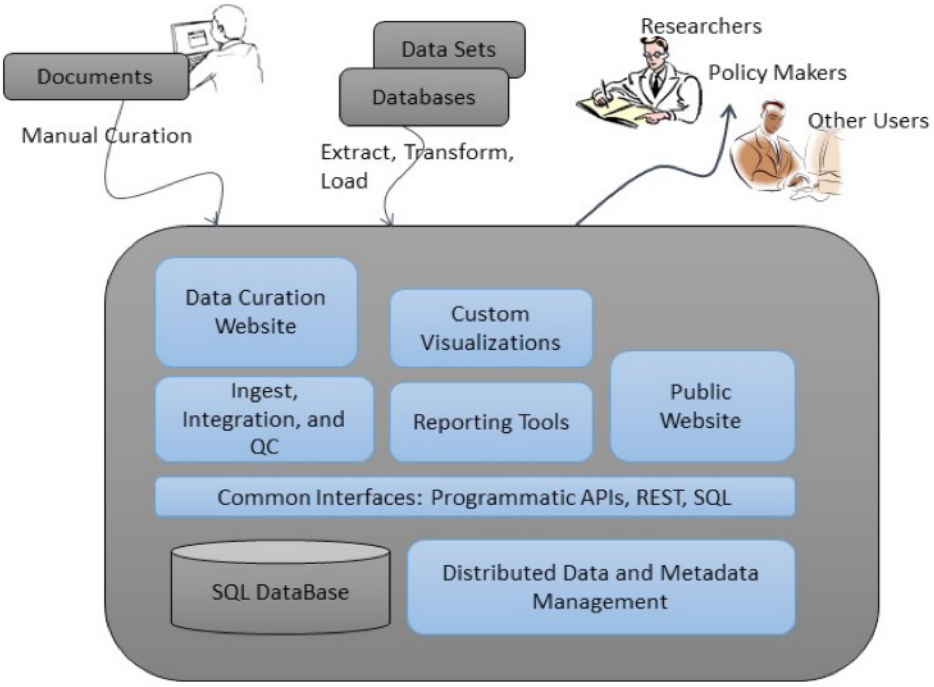
renci

www.renci.org

Abstract

Research on regional entrepreneurial ecosystems suffers from a reliance on outdated data sources and data collection systems that lack innovation. Three main challenges exist: first, research on entrepreneurial regions has traditionally relied on macro-level data sources; second, functional boundaries of entrepreneurship within a region are poorly defined; and third, limited data exists on the individuals who create entrepreneurial firms. Additionally, social factors, regional idiosyncrasies, and dynamic processes influence entrepreneurship but are not typically accounted for in data-driven research. We have focused our research efforts on entrepreneurship within the Research Triangle Park (RTP) region of North Carolina. We have approached our research through detailed analysis of granular, integrated, time-series data on individual companies and the social, geographic, economic, political, and institutional factors that influence entrepreneurship. To enable our efforts, we have created a unique database and data system—*Triangle Genesis Data System*—that contains more than 50 years of data captured from greater than 30 distinct data sources on approximately 4500 companies in the RTP region. Our data system addresses major data science challenges related to data curation, management, integration, visualization, and governance. Herein, we describe our experience with the research, development, deployment, and testing of the Triangle Genesis Data System.

Key words: data system; time-series data; visualization; analytics; social science; regional economics



I. Introduction

Entrepreneurship involves the commercialization of a new idea or creative concept. Yet, research on entrepreneurship often relies on outdated data sources or data collection systems that are anything but innovative. As a result, research on entrepreneurial ventures, especially research on regional entrepreneurial ecosystems, obscures the full range of factors contributing to firm formation and success. Theoretically, the field of entrepreneurial studies has advanced to recognize the essential contribution of multiple regional actors, support programs, and inter-firm interdependences in shaping entrepreneurial opportunity and advancement. For empirical research to stay in step with these theoretical insights, new approaches to data collection and curation are needed. Progress on this front not only holds promise for further entrepreneurial development, but can enable policy-makers and practitioners to identify and target gaps in existing business support, with the goal of supporting and strengthening elements of their regional entrepreneurial ecosystem.

Three main challenges exist with the available data sources and current data collection approaches used by businesses and government; these challenges highlight opportunities for the creation of a more robust and flexible data infrastructure. First, research on entrepreneurial regions has traditionally relied on macro-level data sources such as publicly available company information (e.g., dates of incorporation) and governmental records (e.g., geographical regions as defined by the U.S. Census Bureau). Research in regional economics has long defined *regions* as the constituent organizations within a map-defined geographical space [1]. This narrow focus ignores the vast majority of social factors, regional idiosyncrasies, and dynamic processes that influence entrepreneurship and other outcomes of interest.

Moreover, attempts to overcome this limitation have relied on resource-intensive, manual curation of data from unstructured data sources such as annual reports. While traditional approaches to the study of regional economics remain useful, the wealth of new data sources, including digital news feeds and other social media data, as well as the development of new analytic and visualization techniques, provide an opportunity to create micro-level, time-series data to enable detailed study of the many dynamic processes that influence regional economies.

A second, related challenge stems from poorly defined functional boundaries of entrepreneurship within a region. An illustrative example exists within our own study region in North Carolina, specifically, Research Triangle Park (RTP). RTP was conceived as a 6900-acre campus with a unique postal zip code [2], i.e., a clearly defined geographical space. The RTP region is home to a diverse population of technology firms. In particular, the RTP region has established itself as a hotbed of entrepreneurship in the biotechnology sector [3]. Large multinational companies (e.g., IBM, GlaxoSmithKline, Cisco Systems) and a small group of incubated start-up firms reside within the main RTP campus. However, the majority of RTP's entrepreneurial companies are located in small communities adjacent to the main RTP campus, due to restrictive land use covenants within the main campus. This complicates detailed study of entrepreneurship within the region, particularly when granular geographic data on small surrounding communities are suppressed by governmental entities in order to protect confidentiality. Yet, consideration of a wider region is equally complicated because the State of North Carolina considers RTP as a 13-county planning region that covers a much larger geographical area, extending all the way to the Virginia border. Government-issued aggregated data on the larger RTP region are then difficult to interpret and not representative of the micro-geography of counties more closely connected to activities within RTP. Moreover, RTP's anchoring universities and surrounding governmental laboratories (e.g., National Institute for Environmental Health Sciences, Environmental Protection Agency) influence the functional boundaries of the RTP region.

A third, final challenge relates to limited use of details on the individuals who create entrepreneurial firms, be that founders or influential dealmakers. Individuals are important players in the regional economy but are often studied in isolation from their firms and from the institutional supports within a region. Social networks and social capital are best measured at the individual level—after all, connections are likely to be personal and more meaningfully observed when considering connections between individuals. Firms are started by teams of individual entrepreneurs, with heterogeneous characteristics and backgrounds that affect the outcomes of their firms and their use of external resources. No examination of an entrepreneurial regional economy is complete without consideration of the individuals who create entrepreneurial firms.

Additional factors are known to impact entrepreneurship in regional economies such as RTP, but are rarely accounted for in data-driven research. For instance, transportation routes and modes and land usage agreements exert a major influence on company location [4–5]. Com-

panies often are motivated to locate near other companies with similar markets and access to employees with a defined skill set [6]. Small companies may co-locate in multi-tenant buildings or industrial parks and thus may remain invisible when only aggregated data are available on large administrative units [7]. Moreover, the functional boundaries of a region are dynamic and may expand and contract over time due to serendipitous and largely unpredictable economic, social, and political events [8]. In technology-driven regional economies, such as RTP, companies often offer next-generation, cutting-edge products and services that can be difficult to classify due to a lack of standard terminologies [9].

An additional complexity is the variation in company genesis, with companies originating from research institutions [5,10], as spawns from existing established companies [3,11], or through one-time investments by venture capitalists or governmental entities. Once started, companies remain dynamic and benefit from a range of external programs and support from different organizations. Moreover, firms often close, merge, or are acquired—important events that are rarely considered by researchers who study entrepreneurship [12].

Given these three main challenges in data collection and curation, as well as the additional factors that influence regional entrepreneurship, we set out to build comprehensive data on entrepreneurial firms in order to better understand the changing entrepreneurial development and impact of North Carolina's RTP region. Our singular focus on the RTP region has allowed us to conduct in-depth data collection and analysis of the many regional factors involved in entrepreneurship, with the goal of creating a general model for comprehensive approaches to the study of regional economies around the globe, including other fast-growing innovative U.S. regions such as Austin, Seattle, and San Diego, in addition to early regional pioneers such as Silicon Valley and Boston's Route 128 [13].

Our *Triangle Genesis* project aims to achieve comprehensive insight into the temporal dynamics and multiple features influencing regional entrepreneurship through detailed analysis of granular, integrated, time-series data on individual companies and the many social, geographic, economic, political, and institutional factors that define the functional boundaries of a region. Members of our research team (MPF and NJL) have developed an approach of manual extraction and curation of social science data from multiple structured and unstructured information sources. They term this high-accuracy approach *industry-forensics.* We have coupled this approach with a flexible, robust, web-enabled cyberinfrastructure for data storage and visualization and automation and expansion of our processes and capabilities. This unique database and data system—*the Triangle Genesis Data System*—contains more than 50 years of data captured from greater than 30 distinct data sources on approximately 4500 companies in the RTP region, including company names and focus areas, founders, executives, funding sources, patents, and business-related events such as IPOs, investments, layoffs, and closings. The data system includes a web-based curation system, flexible reporting tools, advanced visualization capabilities, and a public website. This paper describes our experience with the research, development, deployment, and testing of the Triangle Genesis Data System.

II. Materials and Methods

II. A. Data Science Challenges Faced by the Triangle Genesis Project Team

Data collection benefited from earlier efforts by Dr. Bill Little, who began to study technology-oriented entrepreneurial companies in RTP in the 1990s. Those initial data were contained in an early database that contained information on 117 companies. The database has since been expanded by MPF and NJL and currently contains data on greater than 4200 companies. The Feldman-Lowe database focuses on technology-intensive companies, defined as companies that develop products and services in the life sciences, information science and communication, gaming, nanotechnology and other emerging technology sectors. Current primary data sources are listed in Table 1.

Some of these are freely available (e.g., U.S. Food and Drug Administration data), whereas others require a subscription (e.g., Thomasnet.com®). In addition to serving as a data source, the primary data sources also are used to determine whether to include a company in the database, with the criteria being three mentions in the primary sources. Secondary data sources are used to provide detailed information about the company and include company websites, annual reports, newspaper articles, trade magazines, press releases, and social networking sites.

With the growth of the database and the desire to take advantage of today's wealth of big data and advanced analytic and visualization capabilities, the Triangle Genesis research team (led by authors MPF, NJL) realized the need to establish a collaboration with a technical

Primary Data Source	Type of Data Elements	Openly Accessible
Council for Entrepreneurial Development	Membership information, mentoring networks, innovators reports	No
Delphion Patent Data	Patent-related data (application, awards)	No
Innovaro Medical Device Licensing	Medical device licensing agreements	No
Microelectronics Center of North Carolina	Membership information, newspaper articles, annual reports, meeting minutes	No
National Establishment Time Series Database	Longitudinal data on job creation/destruction, sales growth performance, location and mobility patterns, corporate affiliations, etc.	No
National Venture Capital Association	Venture capital financing,	No
North Carolina Biotechnology Center	Description of companies, product development information, clinical trial registrations	No
One N.C. Small Business Program	State of North Carolina small business financing program (SBIR/STTR program)	Yes
Quarterly Census of Employment and Wages, North Carolina	Employment data, salary data	No
Small Business Association	Federal small business financing (SBIR/STTR program and other loans)	Yes
S&P Capital IQ	Global small business information	No
ThomasNet.com	Supplier information, product information	No
U.S. Food and Drug Administration	Drug and medical device approvals	Yes
U.S. Securities and Exchange Commission	IPO filings	Yes

Table 1: Triangle Genesis: primary data sources, type of data elements available through each sources, and availability of each source.

team with expertise in cyberinfrastructure, analytics, and visualization (led by authors OK, ST, CPS). From the outset of the collaboration, the research and technical teams have worked together to conduct a needs assessment and identify the challenges that are hindering our research efforts, as well as potential solutions. The initial needs assessment took place from October 2014 through December 2014. Several specific data science challenges were identified during this period (described below and listed in Table 2):

Data curation: The growth in the number of available data sources created resource challenges related to the research team's industrial forensics approach to curation, which involves manual identification of relevant data sources, extraction of data by a primary reviewer, and thorough fact-checking by a secondary reviewer.

Data integration: The availability of multiple data sources created challenges in data integration due to the wide variability in structure and standardization among the data sources, which range from standardized and vetted governmental databases to completely unstructured social networking data.

Data management: The existing Microsoft Access database and associated software were insufficient in several regards: (1) incapacity to support a growing number of data fields (Access sets a maximum of 255 fields); (2) lack of sophisticated integration tools; (3) incompatibility of the existing system with newer versions of Windows; (4) inability to support simultaneous access by multiple users (Access only supports sequential use by individual users); and (5) insufficient security.

Data visualization: The current data system did not allow for dynamic visual displays of the data by firm, geographic region, and other relevant factors. This limitation was hindering both data analysis and the research team's ability to securely share the data with potential collaborators, research funders, and economic policy makers.

Data governance: Over time, the Triangle Genesis project had become more nationally recognized, and the research team was receiving an increasing number of requests for access to the data. This challenge presented both an opportunity to establish new collaborations and an obstacle, in terms of the development of appropriate policies and secure mechanisms for data sharing.

Data Science Challenge	Solution
Data curation	Automate aspects of the data search and extraction process
Data management	Migrate from Microsoft Access platform to web-enabled SQL platform
Data integration	Migrate from Microsoft Access platform to web--enabled SQL platform; incorporate additional custon software for data integration
Data visualization	Implement a web-based infrastructure to support advanced visual analytics
Data governance	Embed automated, policy-based access and reporting capabilities

Table 2: Triangle Genesis: data science challenges and planned solutions.

II. B. Design, Development, Deployment, and Testing of Triangle Genesis Data System

Owing to the expertise of the technical team and the efficient, regular communication between the technical and research teams, the design, development, deployment, and testing process for the Triangle Genesis Data System proceeded quickly, with few surprising twists. During the two-month needs assessment period, the technical team outlined potential solutions (Table 2). The technical team decided to automate several aspects of the manual data search and curation process, thereby reducing the manpower burden. Data management challenges would be overcome by migrating from a Microsoft Access platform to a web-enabled, Microsoft SQL platform. Of note, the technical team had previously been tasked with this type of migration as a result of their work on other projects and hence had considerable experience (e.g., [14]).

Data integration challenges were addressed through SQL-associated tools and additional custom software that was developed by the technical team. Data reporting challenges would be overcome by establishing levels of access for different users and embedding a web-accessed, policy-based reporting system within the new data system. Data visualization challenges would be overcome by implementing a web-based system to support advanced visual analytics, an area in which the technical team also had vast experience (e.g., [15–16]). The reporting and visualization capabilities would also address issues related to secure data sharing.

With these solutions in mind, the technical team developed and deployed a prototype database and website in December 2014. Internal testing of the prototype continued until live deployment of v1 in October 2015. Both technical (e.g., faulty links) and procedural (e.g., new report requests) issues were identified during the internal testing period. Minor fixes were applied and new features were added to rectify those issues. No major issues were identified, however, and v1 remains in place. The technical team has applied continual updates to the system since live deployment. The updates have primarily involved the addition of new features (e.g., reports) at the request of the research team.

III. Results

III. A. Overview of Triangle Genesis Data System

Figure 1 provides an overview of the Triangle Genesis Data System. As discussed above, a web-enabled Microsoft SQL Relational Database Management System (RDBMS) is used for data management and storage. The database is structured around two main entities: companies and founders. Company data include company name and address, year established, sector and technology, company events (e.g., funding awarded, liquidity events), annual employee counts, and institutional supports (e.g., university affiliation, incubation services). Founder data include founder's full name, education history, work history, and connections to other companies (e.g., founder or co-founder of another company). The company and founder data are linked via a relational schema in order to allow users to explore the rich relationships among entities.

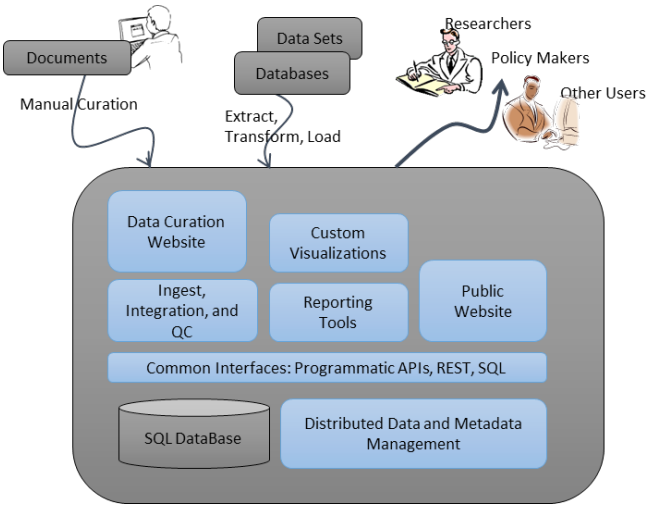


Figure 1. Overview of Triangle Genesis Data System. API = Application Program Interface; REST = Representational State Transfer; SQL = Structured Query Language; QC = Quality Control.

Unlike the Access system, the SQL RDMS is equipped with tools to support data integration. We have supplemented this with custom software to further assist with data integration.

We've coupled the SQL database with iRODS (integrated Rule-Oriented Data System) in order to provide full capabilities for policy-based distributed data and metadata management across Triangle Genesis research team members and collaborators [17–20]. While the data system does not require the use of iRODS, the addition of iRODS provisions several valuable features, including policy-based control of data access, auditing capabilities, and the use of metadata to maintain a provenance trail. These features help to ensure help that only authorized persons gain access to the data at approved levels of granularity (e.g., company and/or founder lev-

el, aggregate level). Of note, iRODS also can support the management of community-contributed data sets, which is something we aim to achieve in the future.

The data system includes semi-automated and automated approaches to data extraction and curation. Manual data extraction and curation remain in place, primarily for unstructured data sources such as newspaper and trade magazine articles. However, all manually extracted data are entered into the system via a Data Curation Website, which was designed to standardize and facilitate data entry. Automated data extraction and curation are used for more structured data sources such as governmental databases. This is accomplished primarily by automated download of pre-selected data elements from within a database or data set and transformation of the data before upload into the database.

We are currently exploring additional automated and semi-automated data extraction approaches such as natural language processing (NLP) and information extraction (IE) techniques. In our preliminary work, we are modifying and testing two approaches to NLP: Named Entity Recognition (NER) and Co-reference Resolution [21]. Our preliminary results suggest that our modified NER approach improves the Stanford NER approach (i.e., the current gold standard for NLP), but this work remains in development and testing.

Usage of the Triangle Genesis Data System has been consistent since the system went live (Figure 2). Authorized users access the data system via password-protected, common user interfaces (e.g., programmatic APIs, etc.). Permission levels are set by the database administrator and correspond to each team member's role (i.e., administrator, researcher, data entry staff, viewer, visualization viewer, report viewer). The system can be customized to enable user authentication via a variety of systems (e.g., Google, Facebook, etc.) A flexible reporting system allows authorized users to generate downloadable reports, and customized, web-based visual analytics provide an extension of the reporting capabilities (these capabilities are described in greater detail below). The data system also is equipped with a public website for dissemination.

Few errors have been encountered since the data system went live (Figure 3). Of the 331 errors reported over the first seven months after deployment, 37.2% involved "data-not-found" errors, and 57.4% were internal server errors. In terms of the data-not-found errors, most were due to invalid link types, links to references or reports that had been removed, or links to inactive reports. The internal server errors were all minor. A common issue involved a duplication of company name, for example, "Glaxo", "GSK", and "GlaxoSmithKline". This issue was resolved by additional data cleaning, allowing for aliases or tags in some cases and manual correction of company names in other cases. Misspellings in the data system itself also were commonly reported in the error reports.

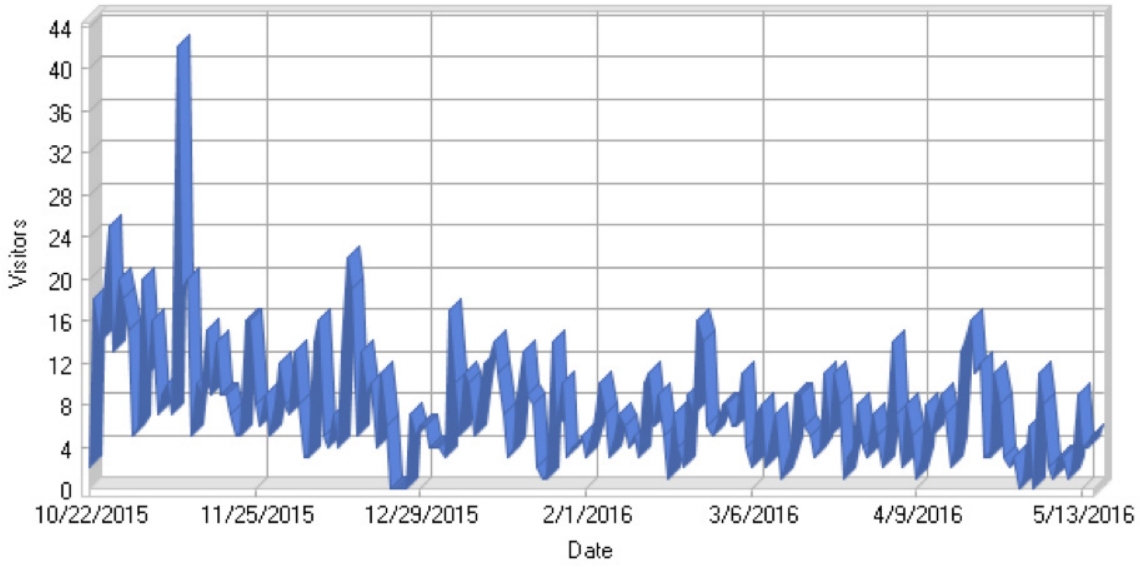


Figure 2. Usage of the Triangle Genesis Data System, from live live dployment in October 2015 through May 2016.

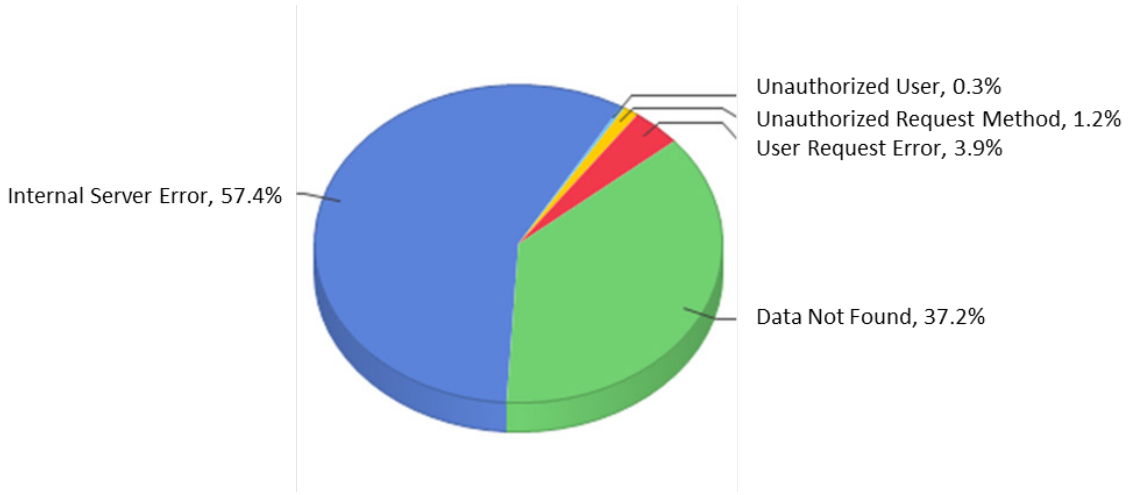


Figure 3. Pie chart showing a breakdown in types of errors encountered with the Triangle Genesis Data System, from live deployment in October 2015 through May 2016.

II. B. Key Outputs of the Triangle Data System

Our Triangle Genesis Data System provides two key outputs that greatly extend the reach of the prior data system: a customizable, policy-based reporting system and web-enabled visual analytics. These are described below.

The reporting system currently supports close to 26 reports that are used by authorized researchers and can be shared, with appropriate permissions and approvals, with research collaborators. The reporting system provides policy-based access to granular, entity-level data that can be downloaded as csv or json files to allow for statistical analysis by authorized researchers. These include reports on: companies by sector and subsector; company affiliation with an institutional or organizational anchor, such as a firm that spun-out of a university or was spawned by former employees of a large corporation; event history by company, sector, and subsector; founders per company; and work history of founders. The reporting system was designed to be flexible in that users can customize new reports without administrator involvement.

Several types of pivot- and geographic-based visualizations are also available to users. The visualizations are designed for data sharing, analysis, and exploration. The visualizations also help with quality assurance of the data (e.g., identification of outliers or missing data points). Figure 4 provides a static view of an otherwise dynamic visualization that demonstrates the remarkable growth in the number of companies in the RTP region, from 1990 through 2012. Figure 5 provides a snapshot of a dynamic visualization that depicts the relationship between company start date and closing date by year, from 1965 through 2015, and with respect to state and national unemployment rates and periods of national recession. Figure 6 shows a static visualization demonstrating networks of founders and co-founders in RTP. Note that only companies with more than one founder are represented in this image; sole proprietors are not included, as the goal was to demonstrate networks.

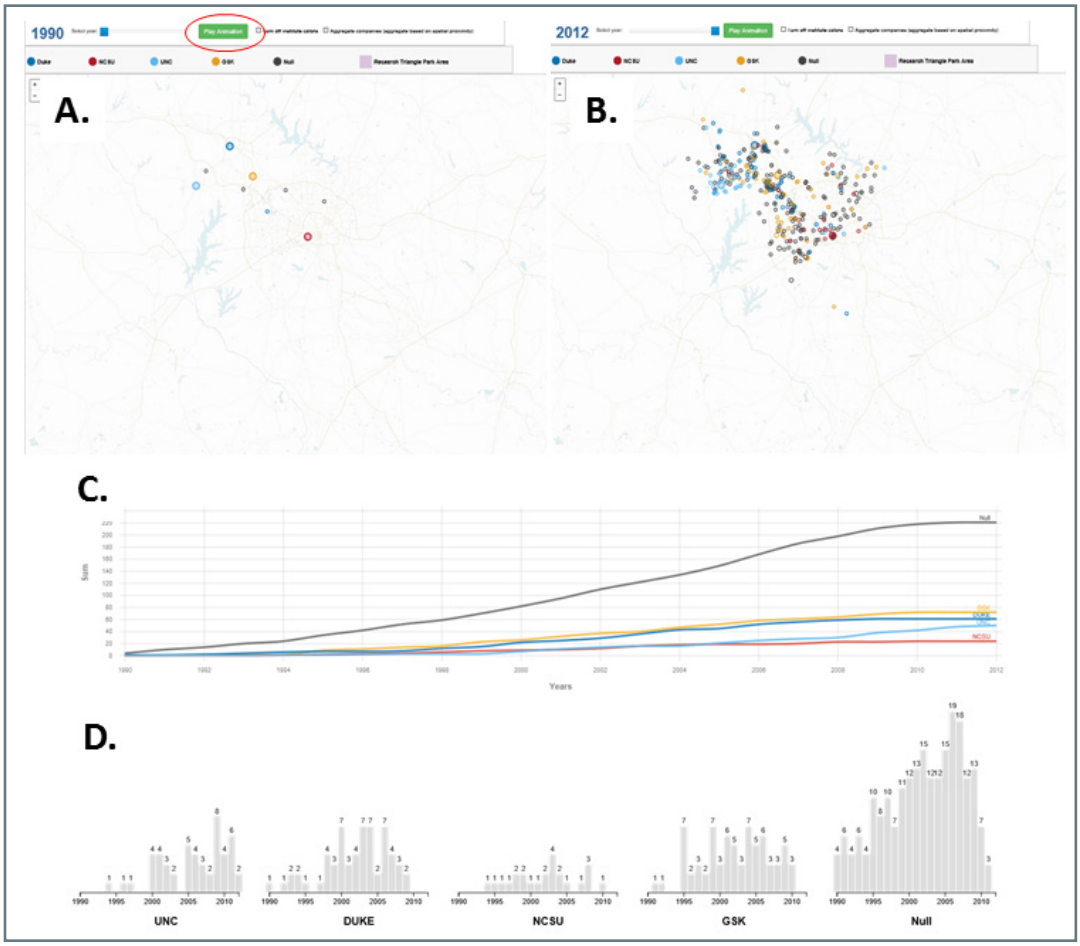


Figure 4. Static snapshot of a dynamic visualization showing the growth in companies within the RTP region from 1990 through 2012, with respect to institutional affiliation and geographical location. By hitting the "Play Animation" button circled in red, the user is provided with a temporal display of existing companies, from 1990 through 2012, color-coded by institutional affiliation (Duke [dark blue], NCSU [red], UNC [light blue], GSK [yellow], Null [black]) and overlaid on a geographical map of the region (in faint light blue). (A) Number of companies in 1990. (B) Number of companies in 2012. (C) Line graphs showing the number of companies by anchoring institution and year. (D) Bar graphs showing the number of companies by anchoring institution and year. GSK = GlaxoSmithKline; NCSU = North Carolina State University; Null = not affiliated with Duke, NCSU, UNC, or GSK; UNC = University of North Carolina at Chapel Hill.

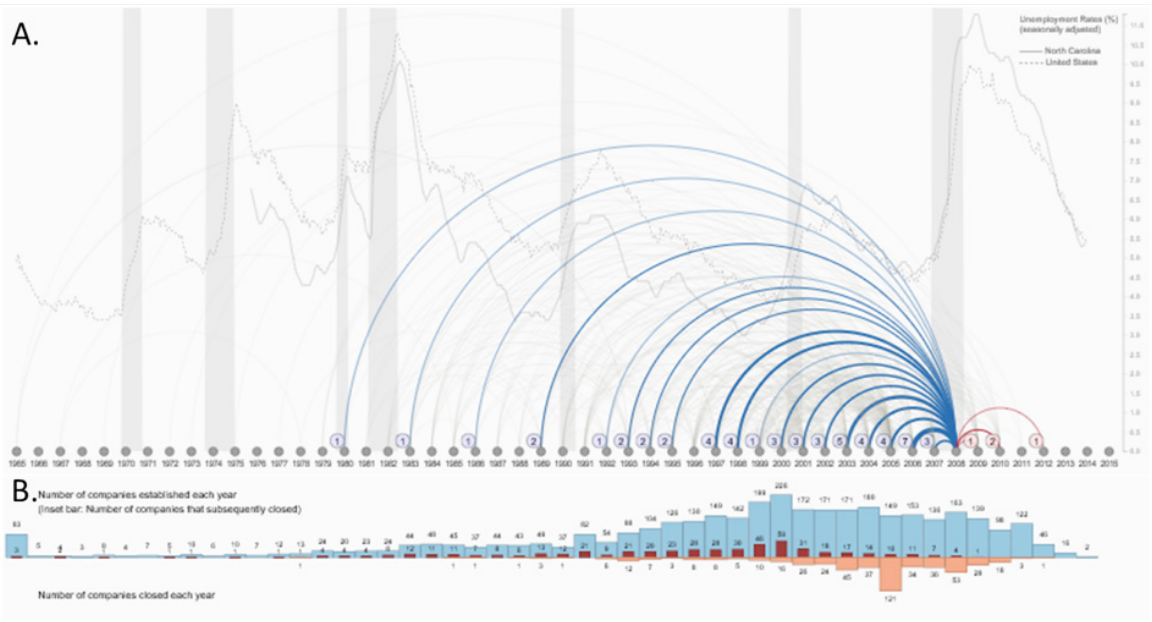


Figure 5. Snapshot of a dynamic visualization showing the relationship between company start date and closing date by year, from 1965 through 2015, with respect to state and national unemployment rates and periods of national recession. (A) In the image shown in this panel, we selected the year 2008 by scrolling a computer mouse over the years (y-axis). The image shows opening dates for companies that closed in 2008 and closing dates for companies that opened in 2008, with respect to unemployment rates for North Carolina (solid line) and the United States (dashed line), as well as periods of recession in the United States (vertical gray bars). (B) This panel shows the number of companies that were established (blue bars) and closed (pink bars) in each year. The red inset bars show companies that were established in a given year but subsequently closed.

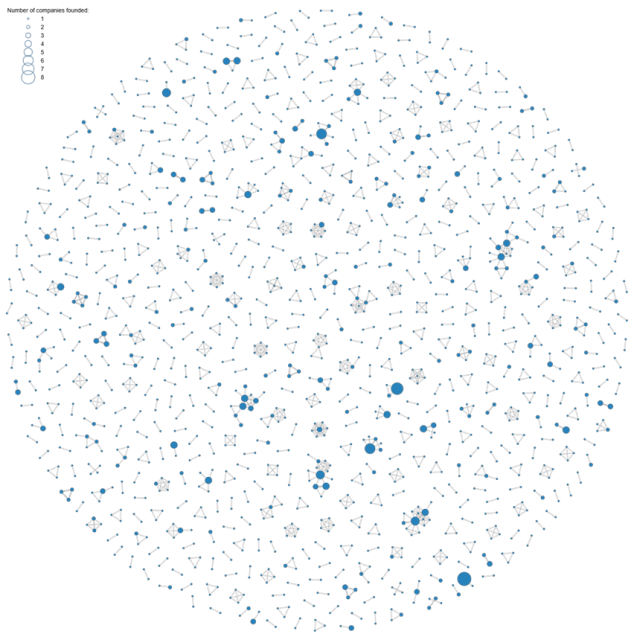


Figure 6. A static visualization demonstrating networks of founders and co-founders in RTP. Nodes represent founders; links represent connections between founders and co-founders. Node size is scaled to represent the number of companies founded (see key in upper left corner). Note that sole proprietors are not represented in the visualization; only companies with more than one founder are displayed, as the goal was to examine networks. The visualization allows one to examine greater details by zooming in on the image.

IV. Discussion

The Triangle Genesis Data System provides a model for research on regional entrepreneurship and the collection, integration, analysis, and secure sharing of granular, time-series data derived from multiple structured and unstructured, public and private, data sources. Our data system overcomes the challenges presented by a traditional reliance on macro-level data sources for research on regional entrepreneurship, poorly defined functional boundaries of entrepreneurship within a region, and the availability of limited data on the individuals who create entrepreneurial firms, whether founders or influential dealmakers, as well as the social factors, regional idiosyncrasies, and dynamic processes that influence entrepreneurship but are not typically accounted for in data-driven research.

Our system also addresses major data science challenges related to data curation, management, integration, visualization, and governance. Specifically, the data system features: a SQL RDMS, coupled with the iRODS policy-based, distributed data and metadata management system; web-based tools to assist and semi-automate the manual data curation process; SQL-associated and custom software tools for data integration; a customizable reporting system; web-based visualization capabilities; and a public website for dissemination. The Triangle Genesis Data System has been supported or used by more than 35 researchers (faculty members, postdoctoral fellows, and graduate and undergraduate students) in numerous fields, including economics, sociology, anthropology, entrepreneurship, land use and regional planning, computer and information systems, journalism, historians, and organizational theorists.

Members of the Triangle Genesis research team are currently using the Triangle Genesis Data System for in-depth research on the dynamic social, geographic, economic, political, and institutional factors that influence entrepreneurship in RTP. Specifically, the data system supports several federally funded or foundation-supported research projects that aim to expand the capabilities of the Triangle Genesis Data System and thereby expand the reach and impact of our research program. Research team members have also submitted proposals for additional funding to improve the data curation and analysis process by incorporating emerging techniques for automated data curation and real-time data analysis. Recent demonstrations indicate that existing automated approaches for information extraction fail short of the precision of human curators [22]. Our goal is to build on our preliminary work on NLP and IE techniques (discussed briefly under Results, [21]) and augment our data system to capture and represent evidence that is uncertain in nature (i.e., assertions), represent that uncertainty in the data and metadata, and enable use of both the data and associated uncertainty for real-time inferences.

The research applications described herein represent just a few of the many ways that the Triangle Genesis Data System currently is being used. For example, our data system serves as a powerful resource for policy development for the North Carolina Department of Commerce, North Carolina Biotechnology Center, Council for Entrepreneurial Development, and a variety of other social and economic policy agencies. We expect many more applications of our system to arise as we continue to grow the rich database and enhance the visual analytic and reporting capabilities of the data system.

IV. A. Limitations

Our Triangle Genesis Data System is in a continual state of research, development, and testing. As such, we are working to overcome several existing limitations.

Although the current system incorporates semi-automated processes for data curation, this process remains largely manual. As noted above, existing automated solutions lack the precision of human curation [22]. We are working to overcome this challenge by improving upon existing NLP and IE solutions [21], as well as factoring in uncertainty regarding assertions on the data.

Another limitation is the time and resources required to match company data from independent sources that do not share a common company identifier. This process is currently accomplished outside of our data system, through a semi-automated process in which we use statistical software to generate a list of potential matches between a new source of company data and the list of companies included in the database. This list is then reviewed by two members of the research team, who independently review the list and additional data sources in order to identify a match (agreed upon by reviewer consensus). A common company identifier is then entered into the data system. We are working to more fully automate this process through the Triangle Genesis Data System.

A final limitation is the time and resource commitment required to develop our data system. The current data system required more than six years of effort by many research and technical team members. While the development of our data system was resource-intensive, we believe that it can serve as a model for other

data systems designed to support research in the social sciences and thus prove to be cost-efficient in the long-term. In this regard, we note that the system is modifiable and flexible enough for adaptation, and we have installations planned at the University of Tennessee at Chattanooga, University of Tennessee at Austin, and several other institutions. The complete architecture and software code for the Triangle Genesis Data System are open source and available upon request.

V. Conclusions

The Triangle Genesis Data System provides a model for a flexible, robust, web-enabled cyberinfrastructure designed to enable in-depth study of the dynamic factors that influence regional entrepreneurship. While developed specifically for research in regional economics, our general approach can be applied to numerous areas within the social sciences for the collection, integration, analysis, and secure sharing of granular, time-series data derived from multiple structured and unstructured data sources. The data system also serves as a powerful resource for social and economic policy makers.

Acknowledgements

This project was supported by RENCI and the National Science Foundation (SMA-1158755, SMA-1262392, SMA-1439532).

Conflicts of Interest

The authors report no conflicts of interest.

How to cite this paper:

Kapeljushink, O., Schmitt, C., Thakur, S., Fecho, K., Feldman, M., & Lowe, N. (2017): Triangle Genesis Data System: Enabling Research on Regional Economics and Innovation. RENCI, University of North Carolina at Chapel Hill. Text. https://doi.org/10.5072/FK2222ZK2H

About the Authors: Maryanne P. Feldman, PhD, is Heninger Distinguished Professor in the Department of Public Policy at UNC. **Nichola J. Lowe, PhD**, is an Associate Professor in the Department of City & Regional Planning at UNC. Feldman and Lowe envisioned the high-level architecture for the Triangle Genesis Data System and identified user needs. **Oleg Kapeljushnik** is a Software Developer at RENCI. **Sidharth Thakur, PhD**, formerly served as Senior Visualization Researcher at RENCI and currently serves as Data Scientist and Visualization Researcher at Intel. **Charles Schmitt, PhD**, formerly was Chief Technology Officer and Director of Informatics at RENCI and now serves as Director of Data Science at the National Institute of Environmental Health Sciences. Kapeljushnik and Thakur designed, implemented, and tested the Triangle Genesis Data System under the scientific direction of Schmitt. **Karamarie Fecho, PhD,** is a biomedical consultant and writer at RENCI. Fecho oversaw the preparation of this technical report.

References*

1. Markusen, A. Studying regions by studying firms. The Professional Geographer. 1994, 46(4), 477-490.
2. Link, A.N. From seed to harvest: the growth of the Research Triangle Park. Research Triangle Foundation of North Carolina: Research Triangle Park, NC, 2002.
3. Feldman, M.P.; Lowe, N.J. Triangulating regional economies: realizing the promise of digital data. Res. Policy. 2015, 44(9), 1785–1793.
4. Audretsch, D.B.; Lehmann, E.E. Warning, S. University spillovers and new firm location. Res. Policy. 2005, 34(7), 1113–1122.
5. Bathelt, H.; Kogler, D.F.; Munro, A.K. A knowledge-based typology of university spin-offs in the context of regional economic development. Technovation. 2010, 30(9), 519–532.
6. Aharonson, B.S.; Baum, J.C.; Feldman, M.P. Desperately seeking spillovers? Increasing returns, industrial organization and the location of new entrants in geographic and technological space. Ind. Corp. Change. 2007, 16, 89–130.
7. Sassen, S. New York City's informal economy. In: Portes, A., Castells, M., Benton, L.A. (Eds.), In: The Informal Economy: Studies in Advanced and Less Developed Countries. The Johns Hopkins University Press: Baltimore, Maryland, 1989.
8. Feldman, M.P.; Graddy Reed, A.; Lanahan, L.; McLaurin, G.; Nelson, K.; Reamer, A. Innovative data sources for regional economic analysis. ebook:Washington, DC, 2012.
9. Feldman, M.P.; Lendel, I. Under the lens: the geography of optical science as an emerging industry. Econom. Geogr. 2010, 86 (2), 147–171.
10. Rothaermel, F.T.; Agung, S.D.; Jiang, L. University entrepreneurship, a taxonomy of the literature. Ind. Corp. Change. 2007, 16(4), 691–791.
11. Klepper, S. Employee startups in high-tech industries. Ind. Corp. Change. 2001, 10(3), 639–674.
12. Lichtenstein, B.B.; Carter, N.M.; Dooley, K.; Gartner, W.B.. Exploring the temporal dynamics of organizational emergence. J. Business Venturing. 2007, 22, 236–261.
13. Saxenian, A. Regional Advantage. Culture and Competition in Silicon Valley and Route 128. First Harvard University Press: Boston, MA, 1996.
14. Henderson, L.M.; Benefield, T.; Marsh, M.W.; Schroeder, B.F.; Durham, D.D.; Yankaskas, B.C.; Bowling, J.M.. The influence of mammographics technologists on radiologists' ability to interpret screening mammograms in community practice. Acad. Radiol. 2015, 22(3), 278–289.
15. Thakur, S. Effective data visualization: what users want. RENCI TriUPA Panel. May 25, 2011. http://slideplayer.com/slide/4674520.
16. Thakur S. Visualization of time, people, and spaces. UNC Libraries Research Hub. November 23, 2015. https://www.youtube.com/watch?v=9q5P9ZsLayQ.
17. Rajasekar, A.; Moore, R.; Hou, C.Y.; Lee, C.A.; Marciano, R.; de Torcy, A.; Wan, M.; Schroeder, W.; Chen, S-Y.; Gilbert, L.; Tooby, P.; Zhu, B. iRODS Primer: integrated Rule-Oriented Data System. Synthesis Lectures on Information Concepts, Retrieval, and Services, 2010a, 2(1), 1–143.
18. Rajasekar, A.; Moore, R.; Wan, M.; Schroeder, W.; Hasan, A. Applying rules as policies for large-scale data sharing. In: Proceedings of the UKSim/AM SS First International Conference on Intelligent Systems, Model-

ling and Simulation (ISMS). IEEE Computer Society Washington: Washington, DC, 2010b, pp. 322–327.

19. Schmitt, C.P.; Wilhelmsen, K.; Krishnamurthy, A.; Ahalt, S.C.; Fecho, K. Security and privacy in the era of big data: iRODS, a technological solution to the challenge of implementing security and privacy policies and procedures. RENCI/National Consortium for Data Science White Paper. RENCI, University of North Carolina at Chapel Hill: Chapel Hill, NC, 2013. http://www.renci.org/wp-content/uploads/2014/02/0313WhitePaper-iRODS.pdf.

20. Fortner, B.; Ahalt, S.; Coposky, J.; Fecho, K.; Heinzel, S.; Krishnamurthy, A.; Moore, R.; Rajasekar, A.; Schmitt, C.P.; Schroeder, W. Control your data: iRODS, the integrated Rule-Oriented Data System. RENCI/iRODS Consortium White Paper. RENCI/iRODS Consortium, University of North Carolina at Chapel Hill: Chapel Hill, NC, 2014. http://renci.org/wp-content/uploads/2014/07/0214WhitePaper-iRODS-2-FINAL-v6.pdf.

21. Wang, Y.; Ma, H.; Lowe, N.; Feldman, M.; Schmitt, C.P. Business event curation: merging human and automated approaches. In: Proceedings of the Thirtieth Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence (AAAI-16). IAAA: Palo Alto, CA, 2016.

22. Wei, C.; Davis, A.; Leaman, R.; Davis, A.P.; Mattingly, C.J.; Li, J.; Wiegers, T.C.; Lu, Z. Overview of the BioCreative V Chemical Disease Relation (CDR) Task. In: Proceedings of the Fifth BioCreative Challenge Evaluation Workshop, 2015. http://www.biocreative.org/media/store/files/2015/BC5CDRoverview.pdf.

*All hyperlinks were last accessed on July 20, 2017.