Data Set Feature

Journal of Conflict Resolution
2019, Vol. 63(6) 1579-1592

DOI: 10.1177/0022002718824635
journals.sagepub.com/home/jcr

NewGene: An Introduction for Users

D. Scott Bennett[1], Paul Poast[2], and Allan C. Stam[3]

Abstract

This article introduces *NewGene*, a complete redesign of the popular *EUGene* software. Like its predecessor, *NewGene* is designed to eliminate many of the difficulties commonly involved in constructing large international relations data sets. *NewGene* is a stand-alone Microsoft Windows and OSx-based program for the construction of annual, monthly, and daily data sets for a variety of *decision-making units* (e.g., countries, leaders, organizations) used in quantitative studies of international relations. It also provides users the ability to construct units of analysis ranging from monads (e.g., country-year), to dyads (e.g., country1-country2-year), to extra-dyadic observations called *k*-ads (e.g., country1-country2-year, . . . , -country*k*-year). *NewGene's* purpose is to provide a highly flexible platform on which users can construct data sets for international relations research using preloaded data or by incorporating their own data. The software is freely available at http://www.newgenesoftware.org/.

Keywords

statistical analysis, data management, software, *EUGene*

[1]Department of Political Science, Pennsylvania State University, University Park, PA, USA
[2]Department of Political Science, University of Chicago, Chicago, IL, USA
[3]Dean, Frank Batten School of Leadership and Public Policy, University of Virginia, Charlottesville, VA, USA

Corresponding Author:
Paul Poast, Department of Political Science, University of Chicago, Pick Hall, 5828 S. University Avenue, Chicago, IL 60637, USA.
Email: paulpoast@uchicago.edu

The software program *EUGene* is a standard tool for quantitative international relations scholars. Initially released in 1998, *EUGene* provides a common software infrastructure for data set construction. It makes routine a set of cumbersome data preparation tasks and keeps track of research design choices. Scholars use the software to create data sets with units of analysis common to quantitative international relations research: the country-year, directed-dyad-year, or directed-dispute-dyad-year. By performing the technical data manipulations necessary to setup data sets, *EUGene* allowed scholars to allocate more time to theorizing and actual data analysis.

While this is all to the good, there is a problem: despite its use as a general-purpose data management software, *EUGene* was not actually created to fulfill that function. Instead, Bennett and Stam (2004) developed *EUGene* primarily to generate data for variables found in Bruce Bueno de Mesquita and colleagues' version of an expected utility theory of war (Bueno de Mesquita 1981, 1985; Bueno de Mesquita and Lalman 1992). This fact did not impede most scholars for most of the past two decades, but times have changed. The data management needs of international relations scholars have advanced since *EUGene*'s initial launch. Three advancements in particular highlight the limits of *EUGene*.

First, more work is moving beyond the state-to-state dyad as the standard unit of analysis. For instance, Fordham and Poast (2016) and Lupu and Poast (2016) use extra-dyadic units of analysis to study the formation of, respectively, defense pacts between states and nonaggression pacts between former rivals. Make no mistake, the state-to-state dyad is still a widely applied unit of analysis in international relations research (Poast 2016). But a key criticism of *EUGene* was that the program led to a convergence on the dyad as the best unit of analysis for international relations research (Bennett 2011). *EUGene* is simply ill-suited for adapting to the current trend of using extra-dyadic units of analysis.

Second, quantitative research in international relations has moved to evaluating alternatives to states, such as leaders or organizations, and to smaller time units, such as months or even days. For instance, Horowitz, Stam, and Ellis (2015) evaluate the war propensity, not of states but of leaders of states. But *EUGene* was oriented toward providing data related to states, typically measured annually. *EUGene* lacks the flexibility to create data sets where states and/or years do not constitute the unit of analysis.

Third, the early part of the twenty-first century has witnessed an explosion in the collection and distribution of data by international relations scholars. Scholars now require a software that can help individual researchers consolidate and organize the data sets he or she requires, work across data sets, and produce data sets for analysis in a principled manner. Instead, *EUGene* was preloaded with a limited set of variables and it was difficult for users to add their own data. But it is now simply not possible to preload a software package with a sufficient number of data sets and variables; the needs of researchers are just too varied and dynamic. Relatedly, *EUGene* was oriented toward conflict scholars who study interstate violence, particularly war.[1] While still of great importance, this rendered *EUGene* less useful for

international relations scholars studying international political economy or conflict scholars interested in exploring civil wars and internal political violence.

Given these trends, we have completely reconceptualized and reprogrammed *EUGene*. Indeed, the changes are so extensive that we refer to the new software as *NewGene*. *NewGene* is a Microsoft Windows and OSx software program designed to minimize many of the difficulties commonly involved in constructing data sets used in international relations and political violence research. *NewGene* is a stand-alone program for the construction of annual, monthly, and daily data sets.

NewGene allows data sets to be constructed around our newly defined concept of *decision-making unit* (DMU). A DMU can be a country, a leader, an organization, an alliance, or any of a variety of other actors studied in quantitative analyses of international relations and conflict. Defining the DMU is completely up to the researcher. Once a DMU is chosen or constructed, *NewGene* provides users the ability to construct units of analysis ranging from monads (e.g., country-year) to dyads (e.g., country1-country2-year) and to extra-dyadic observations called *k*-ads (e.g., country1-country2-year, . . . , country*k*-year).

NewGene's purpose is to provide a highly flexible platform on which users can construct data sets using preloaded data or by incorporating their own data. *NewGene* accomplishes its purpose by automating a variety of tasks necessary to integrate several data building blocks commonly used in testing theories. *NewGene* assembles an output data set based on the user's choices for the unit of analysis, population of cases, and variables to include. The output data set is in comma-delimited .csv format, which can easily be imported into any statistical analysis software program. Users of *NewGene* do not need to be able to write a single line of computer code in order to merge data, read data from input files of varying formats, or convert data into common units of analysis. One need only enter information into menus, either by clicking on the appropriate box or entering text to answer a question, and possibly conduct some basic column or row deletion procedures once the .csv file is created. By guiding users through the process of data management, *NewGene* gives scholars more time and energy to spend on theory development.

This article introduces *NewGene* by describing its functions and operation. We will note its advances over *EUGene* and demonstrate how the two programs compare with an example of constructing a standard state-to-state dyadic data set. *NewGene* is available as freeware at http://www.newgenesoftware.org/. We should note that *EUGene* will continue to be available at http://eugenesoftware.org/, but without regular updates.

Simplicity and Flexibility in Data Management

The first goal of *NewGene* is to simplify data management for applied researchers. An obvious but critically important first step when conducting quantitative data analysis is to create an accurate and appropriate data set. Because this can frequently be an onerous and time-consuming task, *NewGene* was created as a flexible menu-driven tool that allows users to:

- merge data sets,
- choose variables for inclusion in final data sets from a variety of input sources without manually writing code to merge various input data, and
- to provide flexible software that makes it relatively easy for users to incorporate their own data and then merge it with existing data sets.
- Facilitate analysis replication and validation by providing a single program for data set creation that will produce the same results for all users, eliminating the problem of hidden or forgotten steps typically encountered when attempting replication.

As with *EUGene*, the *NewGene* interface provides users with a series of options and tabs. The "Output data set" tabs (visible if one unchecks the "Simple mode" box in the upper left-hand corner) contain the commands and options for choosing variables to include in an output data set, selecting the number of DMU's in each observation of the output data set (e.g., create a data set, i.e., monadic, dyadic, triadic), restricting the units to be included in the output data set (e.g., if the units are countries, this will allow users to select only countries in North and South America), setting the time span of the output data set (e.g., choose the years 1900 through 1930), naming the output data set, and observing—via the status window—*NewGene*'s progress in generating the output data set.[2] The "Input data set" tabs contain the commands and options for importing new data sets, creating new DMUs, inputting a list containing the different members of that new DMU group (e.g., a list of all of the countries in the world), and matching the new DMU to a time unit (e.g., year, month, or day).

Overall, by organizing a number of routines and choices critical to the management of data sets, *NewGene* allows analysts to proceed more quickly to the stage of testing theoretical arguments. The program's flexibility allows it to be used for multiple research agendas, and future updates will allow it to be used in projects that we cannot yet anticipate. The remainder of this section will elaborate on the features offered by *NewGene* and how these are advancements over *EUGene.*

Units of Analysis: From Dyads to k-ads

Perhaps the biggest difference between *EUGene* and *NewGene* is in the unit of analysis. *EUGene* provided users with the choice of three units of analysis: monadic country-years, undirected dyadic country-country-years, and directed dyadic country-country-years. Each remains available in *NewGene*. However, subsequent research revealed that the monad and dyad are inappropriate for studying some key events in the international system, particularly multilateral events (Poast 2010). Multilateral events can be thought of as events that follow a k-adic process (where $k \geq 2$), rather than a purely dy-adic process (where $k = 2$). Using Monte Carlo simulations, Poast (2010) shows that dyadic data cannot recover a k-adic process. Instead, one must use k-adic data, meaning a data set containing all combinations of actors (e.g., actors A, B, and C can form four multiactor combinations: AB, AC, BC, and ABC) up to the

largest sized actor in which the event occurred. For example, if one is analyzing the onset of wars using a data set of 100 countries and the largest war that the scholar is considering contained six countries, then that scholar must consider all combinations of six, five, four, three, and two actors of the 100 country system. Therefore, rather than limiting users to either monads or dyads, *NewGene* provides users the ability to create *k*-ads, where *k* can be any number equal to or greater than one.

The past decade has also witnessed a surge of conflict and IPE research utilizing network analysis (Hafner-Burton, Kahler, and Montgomery 2009; Maoz 2010; Kinne 2013; Cranmer, Menninga, and Mucha 2015; Dorff and Ward 2013; Dorff 2017; Greenhill 2016; Avant and Westerwinter 2016; Chyzh 2016). In light of this development, the flexibility of the underlying code used to create *NewGene* means future releases will be able to format data as adjacency matrixes (i.e., a square-matrix where the cells indicate if two nodes share a connection). Since the creation of this form of data output can prove time consuming, doing so will heighten *NewGene*'s value to network scholars.

Variables: Country Characteristics and Beyond

The heart of *NewGene* (as with *EUGene*) is the ability of users to select variables to include in a data set simply by checking a few boxes. The variables preloaded into *NewGene* come from several important and up-to-date data sets. These include commonly applied data sets from *EUGene* that capture characteristics of states or state-to-state relations. Such data sets include Polity IV democracy scores and ancillary components (Jaggers and Gurr, 1995), Correlates of War (COW) capability data (Singer, Bremer, and Stuckey 1972), and COW Militarized Interstate Dispute (MID) data (Jones, Bremer, and Singer 1996). But *NewGene*, both through preloaded data and an improved ability to accommodate user-created data, also includes a host of variables capturing the new directions of international relations research. The scope of international relations research is moving beyond focusing on characteristics of the state (or of state-to-state relations) and toward evaluating a wider range of units. Some of these units are "higher" than the state level, such as international organizations (Vabulas and Snidal 2013). Other studies consider units that are "lower" than the state level, such as the leader (Chiozza and Goemans 2004).

Each data set in *NewGene* is placed into a "variable group." For example, the Polity IV regime-type data are found in the variable group labeled "03a. Polity Regime Data (Country-Year)" (where the end of the variable group description includes the unit of analysis in parentheses). Placing the variables into "variable groups" helps organize the large number of variables preloaded with *NewGene*. To make clear the extent to which *NewGene* expands the number of available data sets, *EUGene* offered approximately 600 preloaded variables across a range of data sets. However, about 250 of these variables were used only in the computation of expected utility (e.g., the variable *esWsq12* indicated that "Status Quo is expected in equilibrium in directed dyad cc1 cc2, using s [weighted]").[3] Hence, *EUGene* only truly offered users 350 unique and

Table 1. NewGene Version 1.3 Preloaded Data Sets.

Data Set Name	Brief Description of Data Set	Found in *EUGene*?
COW National Capabilities	Capabilities (military and economic) of countries by year (version 3.1)	Yes
COW Major Power Indicator	List and years that countries are considered major powers (version Majors2011)	Yes
Polity IV	Country regime types and characteristics (version 4.0)	Yes
Archigos	Leader characteristics (version 4.1)	No
LEAD	Leader characteristics (version 1.4)	No
COW Contiguity	Measure of distance between countries (version 3.1)	Yes
Capital-to-Capital Distance	Measure of distance between countries (from *EUGene* version 3.204)	Yes
Cshapes Minimum Distance	Measure of distance between countries (version 0.5-1)	No
COW Interstate War	List and characteristics of interstate wars (version 4.0)	Yes
IWD Interstate War	List and characteristics of Interstate Wars (version 1.0)	No
ICB	Militarized crises (version 11)	Yes
ICOW	Territorial and other dispute level data (version 1.1)	No
COW Intrastate War	List and characteristics of civil wars (version 4.0)	No
PRIO Intrastate War	List and characteristics of civil wars (version 4-2015)	No
IPE Data Resource	Various country-level data relevant to study of global economy (version 3.0)	No
ATOP Alliances	List of each country's alliance memberships (version 4.1)	No
COW Alliances	List of each country's alliance memberships (version 4.1)	Yes
Terrorism Incidents	Suicide Attack Database (from Chicago Project on Security and Threats, September 2018 release)	No

Note: Citations for each data set are available at http://www.newgenesoftware.org/. NewGene also contains three "classic" *EUGene* variable groups that include all variables found in *EUGene* (in directed-dyad, undirected-dyad, and country-year format). COW = Correlates of War.

useful variables from which to create data sets for analysis. In contrast, by expanding the types of data, *NewGene* offers users access to nearly 2,000 variables across of range of preloaded data sets. Table 1 reports the data sets included in *NewGene*, along with an indication of whether that data set was found in *EUGene*.

Time of Analysis: Beyond Years

EUGene was programmed to use only annual data, meaning the *year* served as the time increment. This was appropriate, given that the original purpose of *EUGene* was to compute Bueno de Mesquita's expected utility calculations. Computing

expected utility relied on country-level aggregate covariates (military expenditures, trade flows, Polity score) that were only available as annual data.

In contrast, *NewGene* has the ability to produce data sets using daily, monthly, or quarterly time increments. For example, leaders come into office and leave office on specific days. Hence, conducting analysis with a leader unit of observation may require using daily data (e.g., if studying leader tenure, then we want an observation from March 3, 2001 to June 17, 2005, another observation that runs from June 18, 2005 to some other date). Research designs of this sort require observations at specific and regular intervals (daily, monthly, yearly) as well as the capability to merge data of differing time intervals (e.g., leader-day data combined with quarterly economic data or annual election data). Or consider political economy research on financial markets that utilizes daily financial data such as stock and commodity prices. For example, Schneider and Troeger (2006) look at the influence that political developments within three war regions have on global financial markets. They conduct their analysis by considering daily stock market values. The year is simply the wrong unit of time over which to conduct the analysis.

Merging and Data Conversion

One difficulty with building data sets that combine variables is that input data sets may have different units of analysis and come in different formats, requiring conversion at a fundamental level (in addition to simply merging). For example, some key international relations data sets have the country-year as the unit of analysis (e.g., the COW national capability data, Polity data). Other data sets (or data constructions) have the dyad as the unit of analysis, such as the COW contiguity data set. Still other data sets focus on noncountry units, such as data on leaders.

NewGene carries out necessary conversions among the formats, file structures, and the differing units of analysis of these data sets. *NewGene* links together data sets by treating the country as a data set "key." In other words, all of the preloaded data sets in *NewGene* must in some way relate back to a country and have that country identified with a *COW Country Code* (i.e., where the United States has a code of 2, China has a code of 710, Russia has a code of 365). For example, even if a scholar is conducting analysis on leaders, these individuals are leaders of specific countries. The two most widely used data sets on leaders, the *LEAD* data of Horowitz, Ellis, and Stam (2015) and the *Archigos* data set by Goemans, Gleditsch, and Chiozza (2009), include *COW country code* as a variable. To be clear, this does not prevent the user from importing a data set that does not have country variables, but this data set might have difficulty merging with the preloaded data sets (see below).

User Created Data: Easing the Incorporation of New Data

While *NewGene* notably expands the types of preloaded data available to users, we are not able to incorporate all possible data sets. The data available to quantitative

international relations scholars continue to expand, even while existing data sets are regularly updated. This is why a major improvement of *NewGene* over *EUGene* is in the ability of users to specify their own units of analysis and upload their own data. This is another reason why thinking in terms of a DMU is a useful conceptual advance in *NewGene* over *EUGene*. While *EUGene* only treated countries as DMUs, *NewGene* allows consideration of a host of actors as DMUs (e.g., leaders, organizations) and, most importantly, allows the user to specify his or her own DMU. A detailed, step-by-step tutorial for importing data is available on our YouTube channel (https://www.youtube.com/user/NewGeneDataMangement) and via one of our "Quick Start Guides" (http://www.newgenesoftware.org/quick-start-guides/).

Output Format and Use of Other Software

NewGene is not an analysis program but rather a data management utility. Indeed, some users might find *NewGene* to be a useful tool for cataloging and tracking all of the various data sets on their computer. But to conduct analysis, the merged data set created by *NewGene* must be read into and analyzed by other statistical software. Therefore, *NewGene*'s output files are created in a uniform comma-delimited .csv format that can be read into any statistical analysis package. After creating a data set, users can then load the data into the statistics program of their choice.[4] Moreover, to assist in analysis transparency and the replicability of the choices of the user and the procedures conducted by *NewGene*, the software lists the choices and algorithm steps in a *status* window. *NewGene* also produces a ".txt" file (with the same name as the .csv output file) summarizing the choices of the user and the steps taken by the software to produce the new data set.

NewGene or EUGene: A Comparison

NewGene differs from *EUGene* in important ways. As stated, *NewGene* can more easily incorporate a user's own data, generate units of analysis larger than dyads, and produce data sets that do not use countries to construct the unit of analysis. These are clear advantages of *NewGene* over *EUGene*. But the flexibility built into *NewGene* means users familiar with *EUGene* will identify differences in how *NewGene* functions compared to *EUGene*. To illustrate the differences in the two pieces of software, we will describe the basic steps for constructing a standard dyadic MID data set in *NewGene* and how these steps compare to conducting the equivalent steps in *EUGene*. While this example does not demonstrate many of the features that make *NewGene* distinct from (and an improvement over) *EUGene*, the example will directly illustrate advantages (and disadvantages) of using *NewGene* rather than *EUGene*. Further details on using *NewGene* to construct an MID data set are available in one of our Quick Start Guides (http://www.newgenesoftware.org/quick-start-guides/).

Suppose a user wished to create a data set that tested whether certain characteristics of leaders made a state more inclined to initiate an MID against another state. Since leader data are not included with *EUGene* and adding user data is a cumbersome and lengthy process with *EUGene*, such a question is difficult to answer using *EUGene*. With *NewGene*, in contrast, the user can choose from preloaded leader data sets (namely, the *LEAD* and *Archigos* data sets) or add a newly constructed leader data set. Similarly, if a user wishes to use the day as the time granularity or consider triadic data, these tasks can only be performed in *NewGene*.

Consider constructing a data set available in both *EUGene* and *NewGene*: a dyadic data set with the outcome being the onset of an MID between states A and B and the key explanatory variable capturing the *Polity IV* regime types of the two states. There are some similarities in how both *EUGene* and *NewGene* work with such data. For instance, the user will need to choose from a host of different MID-level variables. These include the hostility level, whether the state is an initiator of the dispute or the state is on "side A." Also, when it comes to adding the *Polity* variable to the data set, both *EUGene* and *NewGene* are similar: one simply needs to find the tab with the polity data and select the desired variable.

But important differences arise between the two software platforms. First, *EUGene* offered a unit of analysis specially tailored to working with MID data (either "directed dispute" or "nondirected" dispute data under the "create data set" menu). *NewGene* does not offer a unit of analysis specially tailored to MID data. Instead, MID data are simply one of the many variable groups from which the user can choose. This is necessary to make *NewGene* a flexible software that can handle a host of units of analysis (not solely dyads). The disadvantage is that the preloaded MID data set will not automatically identify the "nonevent" observations, though this can be addressed with procedures discussed in the above mentioned "Quick Start" guide.

Second, while *EUGene* will not ask the user to consider the number of units in a unit of analysis (again, *EUGene* is "hardwired" to work with dyadic data), a user working with *NewGene* will need to choose the number of countries to include in the unit of analysis (by setting the spinner at the top of the screen to "2"). Although this is not a major issue (as "2" is the default value for the spinner), the presence of the spinner could lead to some confusion for users. The key is that setting "2" will produce dyadic data, "3" will produce triadic data, "4" will produce quad-adic data, and so on.

Third, though both *EUGene* and *NewGene* require the user to go to a different tab to set the time range of the data set (the "population of cases" tab in *EUGene*, the "Prepare Run" tab in *NewGene*), a key difference is in the range of possible dates. *NewGene* provides users with the option of specifying a starting and ending *date*, not only a starting and ending *year* (which is the only option available in *EUGene*).

Fourth, the programs differ in the detail they provide regarding program operation. In both *EUGene* and *NewGene*, the user can initiate the data set creation algorithm with a press of a button (the "Ok" button in *EUGene*, the "Generate Output" button in *NewGene*). But while *EUGene* presents a "status bar" with brief

descriptions of the steps, *NewGene* reports the details of the steps (as well as summarizes the choices made by the user) in the "status" box below the "Generate Output" button.

A fifth difference is in how the user prepares the outputted data for statistical analysis. The raw data output produced by either *EUGene* or *NewGene* is in the universally useable .csv format. While this format can be read by virtually any statistical software program, the data require some preparation before it is usable for analysis. *EUGene* and *NewGene* differ in how the user prepares the data.

Suppose the user wishes to use the data in the *Stata* statistical software. For *EUGene*, the preanalysis preparation requires running a *Stata* .do file automatically produced by *EUGene* (if the user checked the "Stata output" box from the "create Command Files" box under the "Files/Format" tab). For *NewGene*, the preanalysis preparation requires the user to load the .csv file (by using the "import" option under the "File" menu in *Stata*) and making some manual adjustments to the data set. To make concrete the manual adjustments, consider the year 1990. MID #3957, which is the Persian Gulf War, is coded as starting in 1990. *EUGene* only automatically reports the Iraq and Kuwait dyad, as it is the initial dyad in the conflict.[5] *NewGene* automatically produces all dyadic combinations of states involved in the conflict, including states on the same side of the conflict (such as the United States–United Kingdom dyad). A user may wish to delete such "same side" dyads before conducting analysis. At the same time, this may not be the case. It obviously depends on the research question being explored by the researcher. Hence, while deleting these "same side" dyads is an extra task for the user, the benefit is that *NewGene* gives the user the ability to choose how to treat these dyads. Moreover, deleting these extra dyads is almost as straightforward as running a *Stata* .do file (as is required by *EUGene*). Assuming the user has loaded the data set into *Stata*, the user can delete the "same side" dyads by specifying that *Stata* drop all observations in which the variable *sidea_1* is equal to *side_2* (i.e., they both have the value of 0 or both have the value 1). This can be done by typing *drop if sidea_1==sideb_2* in the *Stata* "Command" window. Given that both *EUGene* and *NewGene* require the user to perform some basic formatting and manipulation before the output data are ready for analysis, we view the added cost of asking users to perform this command as outweighed by the heightened flexibility offered by *NewGene*.[6]

Conclusion

Our work with *NewGene* is far from finished. We expect updates to *NewGene* to be available on a regular basis, whether to add features or address user-identified bugs. We will continue producing material guiding users through the use of *NewGene* (and make these materials available on the *NewGene* website [http://www.newgenesoftware.org/] and our YouTube channel [https://www.youtube.com/user/NewGeneDataMangement]). As with *EUGene*, *NewGene* will not solve all data set construction problems. But by creating a user-friendly tool to assist in creating and

managing data sets important to international relations scholars, we intend to provide a valuable resource to help the community of scholars interested in understanding international politics.

Some may suggest that by easing quantitative data accessibility, we are bypassing a filtering process that ensures only skilled users will employ these data. This could open the door for poor analysis. Users may create data sets that they do not fully understand, or users may not be forced to gain an appreciation for the nuances of the data they use (Bennett 2011). We believe that this danger is present even without the availability of tools like *NewGene*. It is now fairly easy to download data for replication purposes. Moreover, when using data sets such as the MID data, even experienced users may miss some nuances in the construction of the data (Gibler, Miller, and Little 2016). *NewGene* does not fix these problems, but it does confront users with choices that they *must* make in the course of constructing a data set. In this way, *NewGene*, like *EUGene* before it, is assisting users in making informed choices about the data sets they construct.

We acknowledge that compared to when *EUGene* was first released in 2001, more international relations scholars today are comfortable managing data sets on their own or merging data sets using software such as *Stata* or *R*. But software like *NewGene* is still necessary for a number of reasons. First, there remain a large number of scholars for whom managing data sets in *Stata* or *R* is not an easy process. This is evidenced by the creation of new online data management software for substate data, such as *Growup* (Luc et al. 2015) and **xSub** (Zukov, Davenport, and Kostyuk 2017), along with the continuing citations to the original *EUGene* paper. Second, even for scholars quite comfortable with merging and managing data sets in *Stata* or *R*, a software like *NewGene* is valuable for teaching students statistical methods, as it allows the instructor to spend more time on actual analysis methods and little (or no time) on having students produce data sets for analysis. Third, even when scholars are adept at using *Stata* or *R* to merge monadic or dyadic data sets, moving to extra-dyadic data sets can prove quite challenging. *NewGene* can assist in this task.

Finally, we view *NewGene* and similar software as valuable responses to concerns over transparency and replication in the social sciences. The Data Access and Research Transparency (DA-RT) movement in political science pushes researchers to make their data available, encourages transparency, and seeks to avoid problems of irreproducibility and secrecy in data analysis (Martin and Peterson 2016). By producing fully replicable data sets, with clearly defined inputs, *NewGene* helps researchers striving to work in compliance with DA-RT guidelines. Moreover, the content of the "Status Window" in *NewGene* contains a summary of all options chosen and the algorithmic steps followed by *NewGene*. The content of this window (which is also included in a ".txt" file produced during each run of *NewGene*) can be added to a technical appendix accompanying any article (a demonstration of these procedures is found on the *NewGene* YouTube channel). This is just one of the many reasons that *NewGene* is an important advancement on *EUGene* and will be a useful tool for wide swath of international relations scholars interested in quantitative analysis.

Authors' Note

Author names are alphabetical.

Acknowledgments

We thank Dan Nissenbaum for his outstanding programming work. We also thank the National Science Foundation for financial support.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The National Science Foundation for financial support (Award ID 1059758).

Notes

1. Although perhaps the most commonly used indicator of interstate violence in *EUGene* was the militarized interstate dispute data, which captures levels of conflict that do not meet the threshold of war (Gochman and Maoz 1984).
2. Note that if the user does have the "Simple Mode" box checked, then only the "Select variables" and "Prepare run" tabs are visible. This is the default for NewGene so that initial users of NewGene can focus solely on using the software to generate data sets using the preloaded data.
3. See the "varlabels" excel sheet in the "inputdat" folder available in the main *EUGene* programs folder when downloading the *EUGene* software.
4. In Stata (version 15), this is done using the "insheet" command. In R (version 3.5.2), this is done with the "read.csv" command.
5. Unless the user goes to the "Originator/Joiner Settings" subtab and clicks on "Create Dyads for Originators and Joiners" rather than "Create Dyads only for Originators."
6. Before using the data set, the user may also wish to remove some redundant variables produced by NewGene, such as the multiple variables reporting the year (year_1, year_2). These redundant variables are a by-product of NewGene being programmed to produce data sets with a wide variety of units in the unit of analysis (not just monads or dyads) and working with data sets using a variety of decision-making units (leaders, countries, organizations, etc.). The user can simply delete the extra variables.
7. Elizabeth Martin and Susan Peterson. "Most political scientists will have to change their habits when the new transparency standards start—as of this month." Washington Post Monkey Cage. January 4, 2016. Accessed on August 25, 2018. https://www.washingtonpost.com/news/monkey-cage/wp/2016/01/04/most-political-scientists-will-have-to-change-their-habits-when-the-new-transparency-standards-start-as-of-this-month/?utm_term=.c48efd77fe3d.

References

Avant, Deborah, and Oliver Westerwinter, eds. 2016. *The New Power Politics: Networks and Transnational Security Governance*. Oxford, UK: Oxford University Press.

Bennett, D. Scott. 2011. "Is EUGene a Collective Bad?" *Conflict Management and Peace Science* 28 (4): 315-30.

Bennett, D. Scott, and Allan C. Stam. 2004. *Behavioral Origins of War*. Ann Arbor: University of Michigan Press.

Bueno de Mesquita, Bruce. 1981. *The War Trap*. New Haven, CT: Yale University Press.

Bueno de Mesquita, Bruce. 1985. "The War Trap Revisited." *American Political Science Review* 7:156-77.

Bueno de Mesquita, Bruce, and David Lalman. 1992. *War and Reason: Domestic and International Imperatives*. New Haven, CT: Yale University Press.

Chiozza, G., and H. E. Goemans. 2004. "International Conflict and the Tenure of Leaders: Is War Still Ex Post Inefficient?" *American Journal of Political Science* 48 (3): 604-19.

Chyzh, Olga. 2016. "Dangerous Liaisons: An Endogenous Model of International Trade and Human Rights." *Journal of Peace Research* 53 (3): 409-23.

Cranmer, Skyler J., Elizabeth J. Menninga, and Peter J. Mucha. 2015. "Kantian Fractionalization Predicts the Conflict Propensity of the International System." *Proceedings of the National Academy of Sciences* 112 (38): 11812-16.

Dorff, Cassy. 2017. "Violence, Kinship Networks, and Political Resilience: Evidence from Mexico." *Journal of Peace Research* 54 (4): 558-73.

Dorff, Cassy, and Michael D. Ward. 2013. "Networks, Dyads, and the Social Relations Model." *Political Science Research and Methods* 1 (2): 159-78.

Elizabeth, Martin, and Susan Peterson. "Most political scientists will have to change their habits when the new transparency standards start—as of this month." Washington Post Monkey Cage. January 4, 2016. Accessed on August 25, 2018. https://www.washingtonpost.com/news/monkey-cage/wp/2016/01/04/most-political-scientists-will-have-tochange-their-habits-when-the-new-transparency-standards-start-as-of-this-month/?utm_term¼.c48efd77fe3d.

Fordham, Benjamin, and Paul Poast. 2016. "All Alliances Are Multilateral: Rethinking Alliance Formation." *Journal of Conflict Resolution* 60 (5): 840-65.

Gibler, Douglas, Steven V. Miller, and Erin K. Little. 2016. "An Analysis of the Militarized Interstate Dispute (MID) Dataset, 1816–2001." *International Studies Quarterly* 60 (4): 719-30.

Gochman, Charles S., and Zeev Maoz. 1984. "Militarized Interstate Disputes, 1816–1976: Procedures, Patterns, and Insights." *Journal of Conflict Resolution* 28 (4): 585-616.

Goemans, H. E., K. S. Gleditsch, and G. Chiozza. 2009. "Introducing Archigos: A Dataset of Political Leaders." *Journal of Peace Research* 46 (2): 269-83.

Greenhill, Brian. 2016. *Transmitting Rights: International Organizations and the Diffusion of Human Rights Practices*. Oxford, UK: Oxford University Press.

Hafner-Burton, Emilie M., Miles Kahler, and Alexander H. Montgomery. 2009. "Network Analysis for International Relations." *International Organization* 63 (3): 559-92.

Horowitz, Michael C., Allan C. Stam, and Cali M. Ellis. 2015. *Why Leaders Fight*. Cambridge, MA: Cambridge University Press.

Jaggers, Keith, and Ted Robert Gurr. 1995. "Tracking Democracy's Third Wave with the Polity III Data." *Journal of Peace Research* 32 (November): 469-82.

Jones, Daniel M., Stuart A. Bremer, and J. David Singer. 1996. "Militarized Interstate Disputes, 1816–1992: Rationale, Coding Rules and Empirical Patterns." *Conflict Management and Peace Science* 15:162-213.

Kinne, Brandon J. 2013. "Network Dynamics and the Evolution of International Cooperation." *American Political Science Review* 107 (4): 766-85.

Luc, Girardin, Philipp Hunziker, Lars-Erik Cederman, Nils-Christian Bormann, and Manuel Vogt. 2015. *GROW^{up}—Geographical Research on War, Unified Platform*. Zurich, Switzerland: ETH.

Lupu, Yonatan, and Paul Poast. 2016. "Team of Former Rivals: A Multilateral Theory of Non-aggression Pacts." *Journal of Peace Research* 53 (3): 344-58.

Maoz, Zeev. 2010. *Networks of Nations: The Evolution, Structure, and Impact of International Networks, 1816–2001*, Vol. 32. Cambridge, MA: Cambridge University Press.

Poast, Paul. 2010. "(Mis) Using Dyadic Data to Analyze Multilateral Events." *Political Analysis* 18 (4): 403-25.

Poast, Paul. 2016. "Dyads Are Dead, Long Live Dyads! The Limits of Dyadic Designs in International Relations Research." *International Studies Quarterly* 60 (2): 369-37.

Schneider, Gerald, and Vera E. Troeger. 2006. "War and the World Economy Stock Market Reactions to International Conflicts." *Journal of Conflict Resolution* 50 (5): 623-45.

Singer, J. David, Stuart Bremer, and John Stuckey. 1972. "Capability Distribution, Uncertainty, and War, 1820–1965." In *Peace, War and Numbers*, edited by Bruce Russett, 19-48. Beverly Hills, CA: Sage.

Vabulas, Felicity, and Duncan Snidal. 2013. "Organization without Delegation: Informal Intergovernmental Organizations (IIGOs) and the Spectrum of Intergovernmental Arrangements." *The Review of International Organizations* 8 (2): 193-220.

Zukov, Yuri M., Christian Davenport, and Nadiya Kostyuk. 2017. "Introducing xSub: A New Portal for Cross-national Data on Sub-national Violence." Working Paper.