The PATH Project: Partnerships for Automated Transformations of Heterogeneous Datasets

Kara Nance
Department of Computer Science
University of Alaska Fairbanks
klnance@alaska.edu

Uma Bhatt
Department of Atmospheric Sciences
University of Alaska Fairbanks
usbhatt@alaska.edu

Brian Hay
Department of Computer Science
University of Alaska Fairbanks
brian.hay@alaska.edu

Abstract

Researchers, analysts, policy makers, and members of the public are all interested in using data for various purposes, but they often do not have the time nor the technological skills necessary to assemble and integrate data from multiple heterogeneous data sources to form revolutionary knowledge discovery products. Building yet another database to try to address this may not be the best approach to solving this problem which requires a major shift in the paradigm. The PATH Project team is working to develop revolutionary service-oriented methods that can provide automated solutions to this problem, by involving multidisciplinary teams with extensive experience in providing computer-science solutions to data-related scientific problems. The resulting service-based suite of tool, addresses an important need for data consumers across a wide range of digital domains.

1. Introduction

The ever-increasing rate at which we collect and store data greatly overwhelms our current ability to make these valuable data resources easily available to a wide range of data consumers. At one extreme, these consumers include scientific researchers, whose increasingly complex models manipulate larger, richer, and more finely-grained datasets gathered from multiple heterogeneous sources, providing the opportunity to investigate phenomena in much greater depth that has ever been previously possible. At the other end of the spectrum are policy-makers and members of the general public, who may be interested in data more closely related to their personal interests, such as the level of contaminants in animal species in their state, or studies describing air quality for a particular city. Despite different levels of technical knowledge, this continuum of data consumers faces similar problems, namely how to assemble and integrate heterogeneous domain-specific data into formats that meets their needs.

For the technically savvy data consumer, this task is possible today, but often only with significant and time-consuming manual preprocessing efforts that could be better spent on scientific analysis. However, the typical data consumer usually only possesses some, if any, of the knowledge necessary to convert relevant data for integration into a useful product. The ability to view products based on multiple heterogeneous datasets in new and novel manners is often the key to enhancing the global knowledge base, making informed decisions, and advancing scientific understanding. Unfortunately, this opportunity is frequently unavailable to data consumers, due to insufficient data access resources, lack of advanced data processing skills, or time constraints.

Clearly the current situation demonstrates an ineffective use of our vast and ever-expanding data resources, and of the time and efforts of data consumers and scientists. As a result, while data that could be used to answer important questions often exists, it remains beyond the reach of many data consumers, and related questions remain partially or fully unresolved. Furthermore, data collected at great cost for one project or purpose often becomes less useful over time, not due to data quality issues or a lack of interest in the topic, but simply because other potential data consumers often do not know it exists, and could not access it in a useful manner if they did. A solution must be found to automate much of this process, and to allow users of a system to easily leverage the work done by others. Such a solution will allow for increased productivity on the part of scientists and researchers, greater data access for all levels of data consumers, and increased throughput and utilization for models and analysis tools.

The Partnerships for Automatic Transformations of Heterogeneous Datasets (PATH) project evolved from the need to advance research and contribute to

978-0-7695-3869-3/10 $26.00

the current state of knowledge in the development of web-enabled systems to integrate heterogeneous datasets. The project builds on proven technologies that allow data consumers to easily complete the task of integrating heterogeneous data from distributed sources. The resulting web-based platform-independent open-source toolkit will allow users to transform the resulting data to form an integrated product that better meets their specific needs. In addition, it facilitates the rapid integration of new datasets into the system with a minimum amount of user interaction, by leveraging the experience gained by the system in dealing with and transforming previous datasets, with guidance from the user when necessary. The ultimate PATH vision is to develop discipline-specific libraries of transformations in order to facilitate future knowledge discovery. This paper uses the Atmospheric Sciences domain as a primary demonstration platform and then extends the transformations utilized in the example to an alternative domain.

2. Background

This general service-oriented system builds on two important projects developed by this research team. In the 1997 meeting of the Arctic Monitoring and Assessment Programme (AMAP) participants from around the world quickly noticed the absence of U.S. environmental data. The justification for this omission was that the AMAP submission process and required formats were too complex and time-consuming for the scientists. In response to this issue, the SynCon Project staff at the University of Alaska Fairbanks worked with AMAP and the U.S. scientists to help resolve this issue. In an initial step, scientists involved were asked what was needed to encourage them to submit their data to a central facility. Their requirements included the following: a) Accept my data in its current format; b) Don't make me do much extra work; c) Don't make me hire a data specialist.; d) Don't ask me too many questions; e) Make my data available in the formats I want and will want everyday, forever; f) Protect my data from unauthorized access.

The SynCon development team successfully met this challenge, by producing intelligent agents that accept data diverse formats as input and translate the data into output formats that met the needs of the AMAP Thematic Data Centres (TDC) [1, 2]. As a result, scientists could concentrate on "science" and the data manipulation process, which would have required hours, days, or even weeks of effort for the scientists to complete, was accomplished extremely efficiently through large-scale automation. The SynCon Project was selected to serve as two of the five Thematic Data Centres (TDC) for the Arctic Monitoring and Assessment Programme. It has become apparent that scientists from around the world are choosing to contribute their data to SynCon even when other data centers were their original targets. In a letter commending the SynCon Project, the AMAP Executive Secretary Lars-Otto Reierson observed that

> "[T]he SynCon solution has proven so successful that part of the scientific community that are responsible for reporting data to the AMAP marine and atmospheric TDCs have preferred to send their data to UAF for incorporation in the SynCon Database instead. This has particularly been the case with the North American scientific community that has no previous experience of reporting to the European-based marine and atmospheric TDCs. UAF is therefore now exceeding it original commitment to AMAP by processing all datasets received." [3]

A second foundational project developed by this research team is the a prototype data source registry has been developed for NASA (NASA award AIST-02-0135) that tracks the structure and access methods for multiple data sources. In addition, it allows associations between components of the data sources to be defined, so that, for example, datasets which contain related data from different sources, each with their own structures, can automatically be integrated into a seamless view. Such associations can be created manually by a user or administrator, but are also created automatically as new data sources are made available, using techniques built on previous work. This allows even a novice user to add a new data source to the registry and have it be available for use in the system with minimal user intervention.

While these foundation products represent positive steps to provide valuable service-based functionality to access important data resources, a more comprehensive approach is required to meet the needs of the broader community of data consumers.

3. PATH Model

Integrating multiple heterogeneous data sources today usually involves a time consuming effort by a user with significant technical knowledge. The proposed service-oriented solution to this problem is a general-purpose open-source toolkit based on

research that combines and builds on the proven technologies of existing foundational components, providing a wide-range of data users with the ability to easily integrate data from multiple heterogeneous data sources.

While the SynCon Project provides a basis for this effort, it leaves open several problems that must be solved in order to allow data consumers at all technical levels to access integrated data products. For example, some approaches involving mobile-agents [4, 5] have shown promise, but such approaches require the data source host to provide a suitable environment in which the agent can operate.

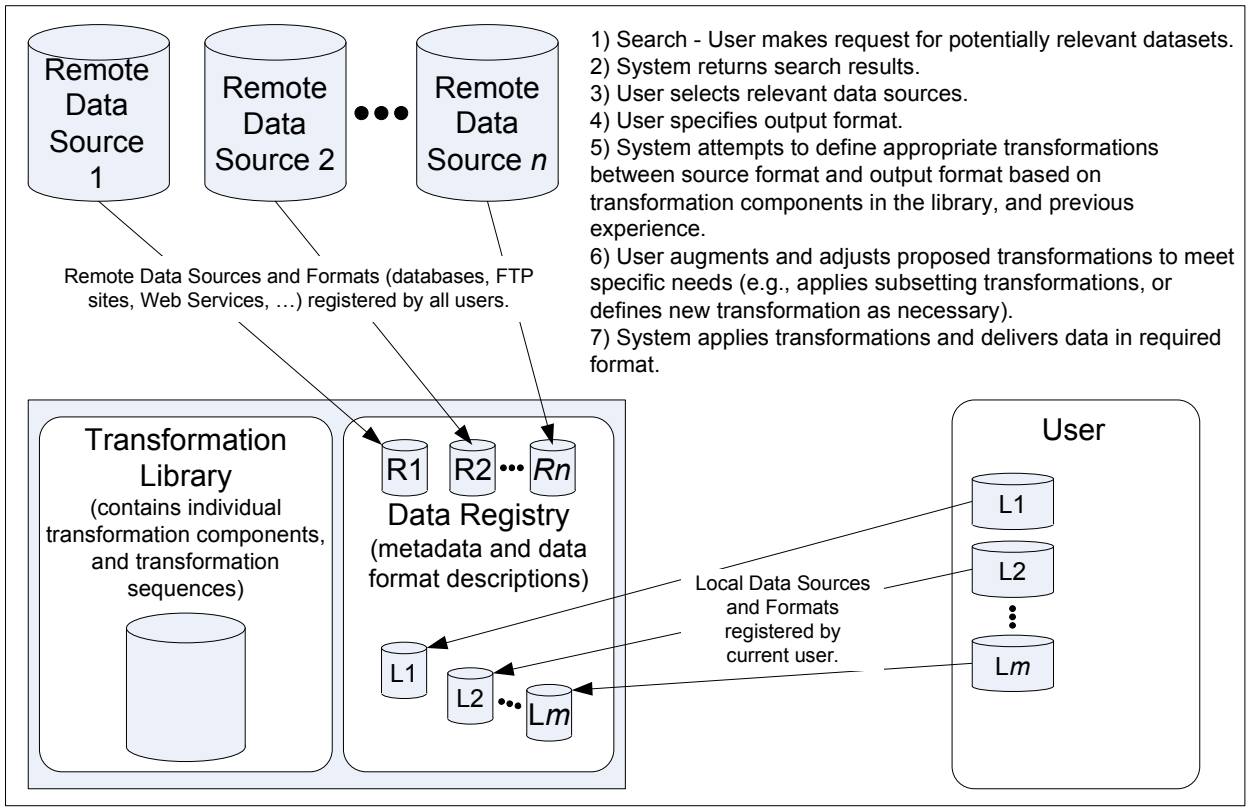

Figure 1: High Level System Design

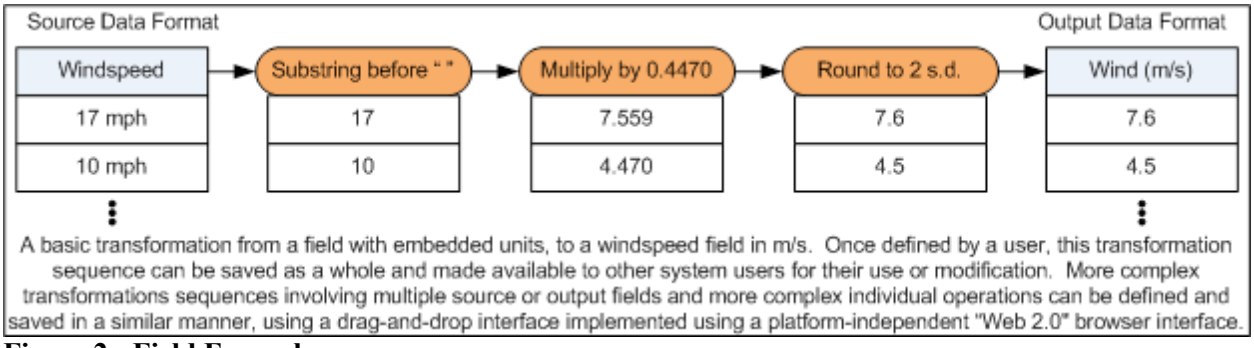

Figure 2: Field Example

The discussion of existing peer-to-peer data integration systems and ongoing research efforts identified in [6] demonstrates other building blocks that are foundational to this research effort and specifically identifies anonymity, security, and access control problems as open and challenging issues in this area.

When a researcher identifies a potential dataset that could provide new scientific knowledge when combined with an existing dataset, the new data

returned will meet one of three conditions. The easiest case is the one in which all of the data is already in the format required by the user or application. Data that is not in the required format, but has a previously defined conversion mechanism is slightly more complex to manipulate. Several of these commonly used conversion mechanisms already exist, such as from netCDF to HDF5, or between the various emerging XML-based standards. The third possibility involves the case in which data requires a conversion, but no defined conversion exists. This is the most complex case, but it can be simplified considerably from the perspective of the user through the use of automated intelligent agents to develop a proposed conversion process, which is then presented to the user for either approval or modification.

A strength of this approach lies in the ability of the system to "save" this conversion process for automatic use with similar datasets in the future. For the PATH Project, this component builds on the SIMON intelligent agent that was developed by this team as part of the SynCon project, and which is in use today to allow data owners to easily submit datasets to the SynCon datacenter in a wide range of undefined formats. While SIMON allows for many input formats, it is focused on specific AMAP TDC output formats, so the PATH Project expands on that base to provide a Transformation Manager that is capable of flexibility with regard to both the input and output formats. Since this component uses a pluggable module approach, new transformations or transformation sequences that are defined, (either automatically by the Transformation Manager, or manually by the project team), can be easily added and re-used for future datasets. It is anticipated that new transformations for commonly requested formats will be developed by user communities, increasing the number of datasets that fit into the second category. Process and field examples are shown in figures 1 and 2.

While several PATH domains have been defined, the following examples demonstrate applications of the PATH Project approach to the Atmospheric Sciences domain and an extension application to an economic domain. Atmospheric Science was chosen for prototyping as it is highly data driven, typically involving large heterogeneous data sets and complex models. Several other application domains, beyond the demonstrated extension domain have been identified for future work, including digital forensics, computer security, and sensor networks.

4. PATH Examples

The following examples were chosen to demonstrate application of the PATH Model to large datasets and subsequent extension to seemingly unrelated datasets that can use the same transformations. Although the target of the prototype test and demonstration environments described herein are related to climate research and economic development, it only serves as a complex example domain. Many additional application and test environments have been identified through international partnerships, ensuring that this innovative research is available to help bridge the digital divide across a broad range of domains and cultures.

4.1 Atmospheric Sciences Example

In climate dynamics research, heterogeneous data sets are investigated in many ways to develop an understanding of the physical processes associated with climate variability. What essentially starts out as an exploration to identify key relationships between climate parameters culminates in a clearer understanding of the important processes that control a given type of climate phenomena (e.g., in Alaska, warmer-than-normal winter temperatures are linked to warm Eastern Equatorial Pacific sea surface temperatures). Software packages (e.g., Matlab or SAS) are generally not useful for this class of investigations because each calculation has slightly different requirements or data formats and these packages often have difficulty with extremely large data sets. As a result, most climate scientists develop new programs for each calculation. Since these calculations are repetitive, the climate scientist will be able to reuse old programs (typically written in FORTRAN or C++) repeatedly. However, these programs will require I/O or parameters changes, which can introduce the possibility of making mistakes. Hence, this has to be done carefully and typically one spends valuable time debugging and checking to make sure the changes are implemented correctly. In addition, a significant amount of time is spent redeveloping the same or similar special purpose programs. A streamlining of this process would be beneficial for hastening scientific discovery for the expert, who understands the calculations, as well as the novice who treats these calculations as black boxes but is interested in the final result. This problem really demands a more general framework that will reduce duplication of effort.

The two examples that are outlined below are simplified versions of what has recently been encountered in research on climate variability in the Arctic as researchers work to answer important

scientific questions. This general class of problem is frequently encountered in climate analysis and is solved at great time expense.

Research Question: *Is Interior Alaska river breakup linked to previous winter severity?*

We began with the hypothesis that if the winter is cooler than normal in Interior Alaska then river breakup will be delayed. To quantify this relationship observational data at stations in Alaska were combined for the months of Dec-Jan-Feb (DJF) for each winter season to construct a time series of minimum air temperature. This data set was obtained from the Alaska Climate Research Center (http://climate.gi.alaska.edu/) and was in tabular ASCII format in degrees Fahrenheit. It was converted to .cvs format and read into a FORTRAN program to construct DJF averages. The river break up data was obtained from the Alaska-Pacific River Forecast Center housed under the National Weather Service (http://aprfc.arh.noaa.gov/data/breakup.php) and is in units of day of the year (Julian Day). The river breakup time series is the average of three stations in Interior Alaska (Bethel, Nenana and Dawson) that are highly correlated. This averaging of the three time series would be an additional transformation, but will be ignored in this example for the sake of simplicity.

The first transformation (T1 in figure 5) that is made to the station data is conversion to degrees Centigrade. The subsequent transformation (T3) constructs a seasonal average, the first time through this exercise we averaged from December to February based on our hypothesis. The time series are then linearly detrended before calculating the correlations since the trend can be quite dominant in climate data, leading to an incorrect interpretation of the results. The next transformation (T4) is to calculate a correlation coefficient between the river break-up time series and seasonally averaged minimum temperature (T2) at numerous stations in Alaska (e.g., Anchorage, Fairbanks, Barrow, Bethel, Homer, Juneau, and King Salmon), which is followed by a transformation (T5) to assess the statistical significance of the correlation using at t-test, which can be an algorithm or a look-up table. The final transformation (T6) is to construct a plot showing the correlations and associated statistical significance (* in the figures indicates significance at the 95% or greater level) at each geographic location of the station data. Next the climate scientist interprets the correlation plot (Figure 3) and would conclude that the relationship between winter temperatures and break-up is not strong. So a subsequent question could be "*Are springtime temperatures more strongly correlated with river ice breakup*?"

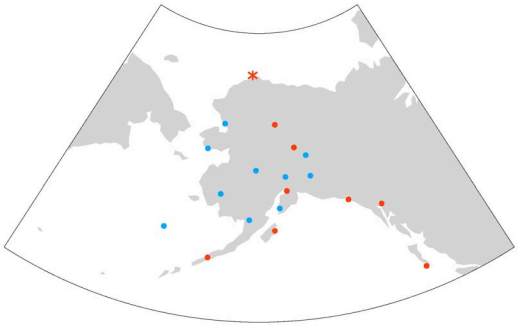

Figure 3: Breakup Correlated with DJF Minimum Air Temperature

The climate scientist would return to the FORTRAN code that read in the station data, modify the code that was used to construct the DJF average to temperature and change the averaging period to March-April-May (MAM), which could be a minor or major modification depending on the original approach. The MAM time series would be correlated with the river breakup and the results plotted to produce Figure 4. The climate scientist would conclude that the spring temperatures are more important in determining the timing of river breakup in Interior Alaska, since more stations display a significant correlation between river breakup and minimum temperature. This also suggests that there is not a long lead time to predict breakup based on minimum air temperatures. The scientist would go on from there and attempt to investigate other variables (e.g., snow depth or precipitation) that could lead to better predictability of river breakup. The effort to perform several similar investigations are shown in figure 5

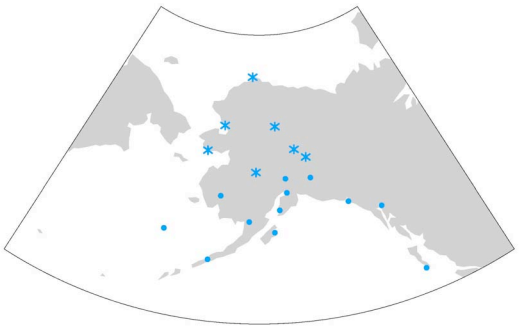

Figure 4: Breakup Correlated with MAM Minimum Air Temperature.

This relatively simple example illustrates the type of analysis that is repeated in climate research. If better tools could speed up the calculation process then the time between asking the question and interpreting the plot would be shortened, enhancing scientific discovery and avoiding many repetitive steps. One could argue that it is easy to write a UNIX script to make plots of every possible combination of correlations with the push of a few buttons using a fairly generic program. The problem with this methodology with regard to climate research is that this large amount of information is not easy to synthesize due to the complex nature of the interactions between the different components of the climate system. It is more productive to make informed choices and judiciously perform new calculations and construct new plots based on intelligent interpretation of results. This process proceeds in a more logical manner and provides more time spent thinking about the underlying physical processes. To facilitate this approach, the transformation sequence, or workflow, used to perform this investigation could be stored as a new transformation, which the scientist could then easily apply to the source inputs (i.e., the river breakup and air temperature time series files of their choice).

Thus from the user's perspective the application of the same transformation sequence to different source data requires minimal effort, allowing them to focus on the science rather than the data management.

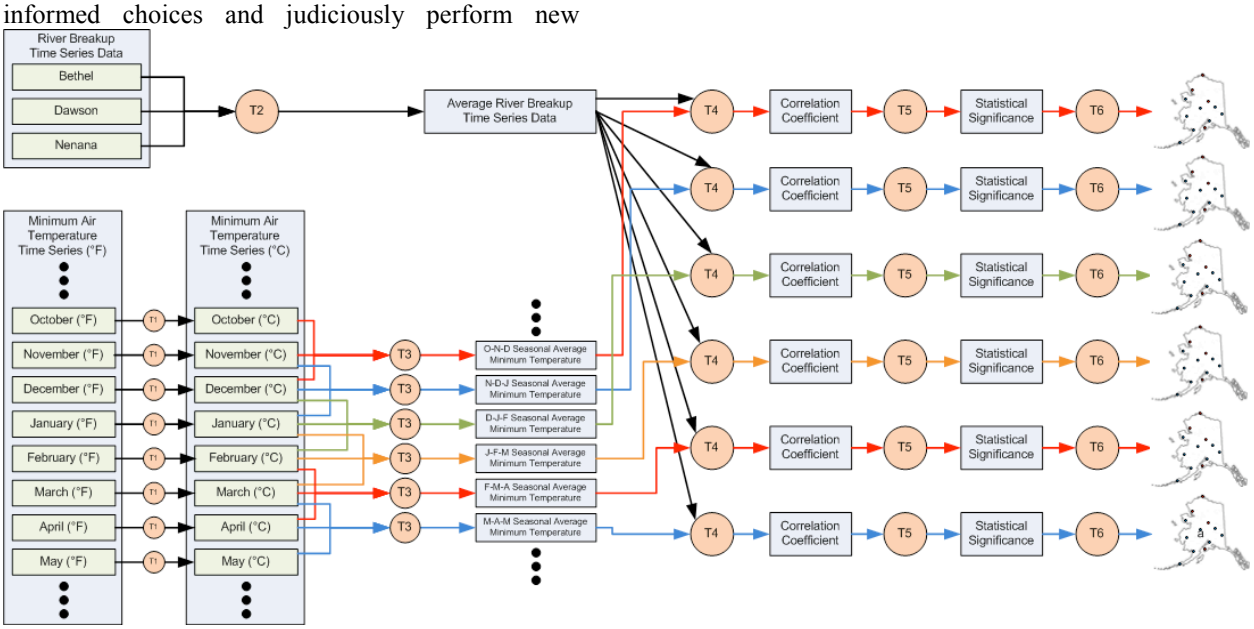

Figure 5: The process used to process River Breakup and Minimum Air Temperature Time Series Datasets into graphic depictions of the correlations.

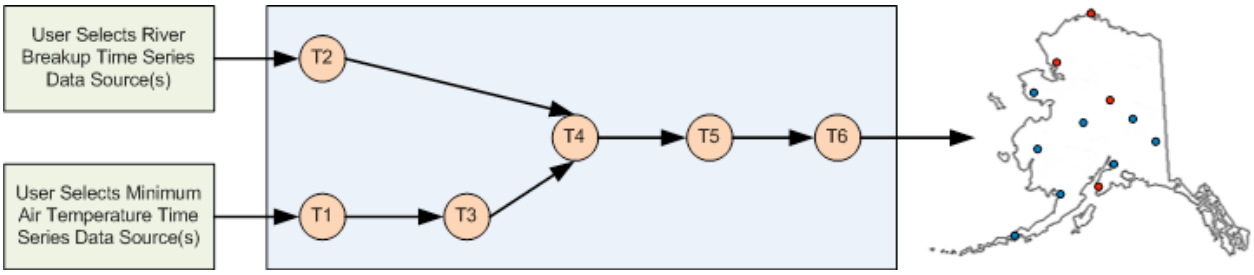

Figure 6 –The transformation sequence applied to user selected source data in Figure 5.

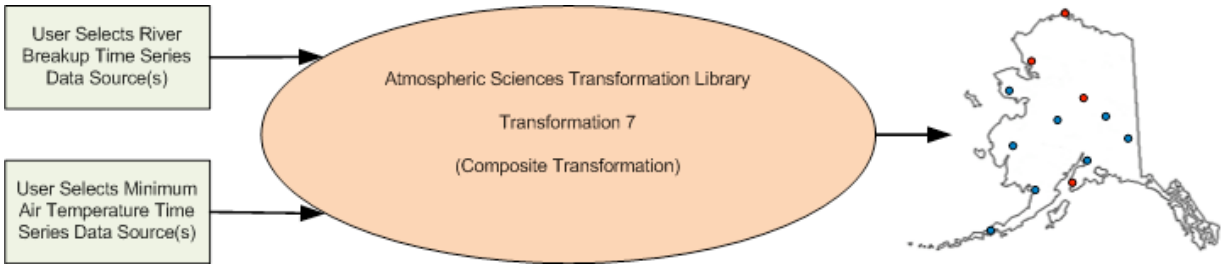

Figure 7 – The application of a single composite transformation to achieve the same result as shown in Figure 6.

Moreover, if the transformation sequence itself is stored as a new transformation, the process from the user's perspective becomes that of simply applying one transformation (which encompasses the entire transformation sequence) to the chosen source data (as shown in figure 7). The type of analysis done in Example 1 has been conducted for Interior Alaska river discharge and ice thickness by our research group. The relationships with the large-scale climate and the different river parameters vary so a range of seasons had to be explored, requiring recoding of the programs with each new data set. Our proposed transformation tools would greatly reduce the time that it took us to proceed from our science question to our final interpretation of the results.

Figure 5 shows an example of the preceding analysis with potential transformations designated by circles on the diagram. (While the individual components are difficult to discern, the overall number and complexity of the transformations is readily apparent in this figure.) Figure 6 shows the transformation sequence being applied to source data selected by the user as described in the previous paragraphs, and figure 7 shows how the composite transformation, which is now stored in the transformation library in addition to the individual transformations, can be applied by a user as a single transformation to user selected source data.

4.2 Economic Example

Many of the transformations used in the previous example could also be applied to other sources of data, both within and beyond the Atmospheric Science domain. One example is that of unemployment data, for which we may want to examine 3 month trends in various regions, many of which experience variable employment rates throughout a given year due to seasonal activities such as fishing and tourism. For a researcher interested in the unemployment trends in these areas, a comparison of three-month averages over a number of years may be valuable. Transformations T3 and T6 from example 1, could be re-applied to this task, in addition to a new transformation (T8) which produces trends from several input data files. This sequence of operations and an example of the resulting graphical output is shown in figure 8. Note that the choice of visualization is determined based on the user population with unique choices for representing increases and decreases in the data as well as different map projections.

A more complex example may involve attempting to correlate the current trends in economic data for a given location to the data for regions which typically lead the economy into and out of recessions or economic booms. For example, it may useful to determine how far Arizona is behind California in entering and exiting periods of growth in order to more effectively predict economic conditions in Arizona. This process may be quite similar to the previous atmospheric science example in which correlations between temperatures and river breakup are being sought, and as such it is quite possible that the entire transformation sequence (i.e., a more complex composite transformation) "Atmospheric Sciences Transformation Library Transformation 7" could be reused for this purpose with little or no modification.

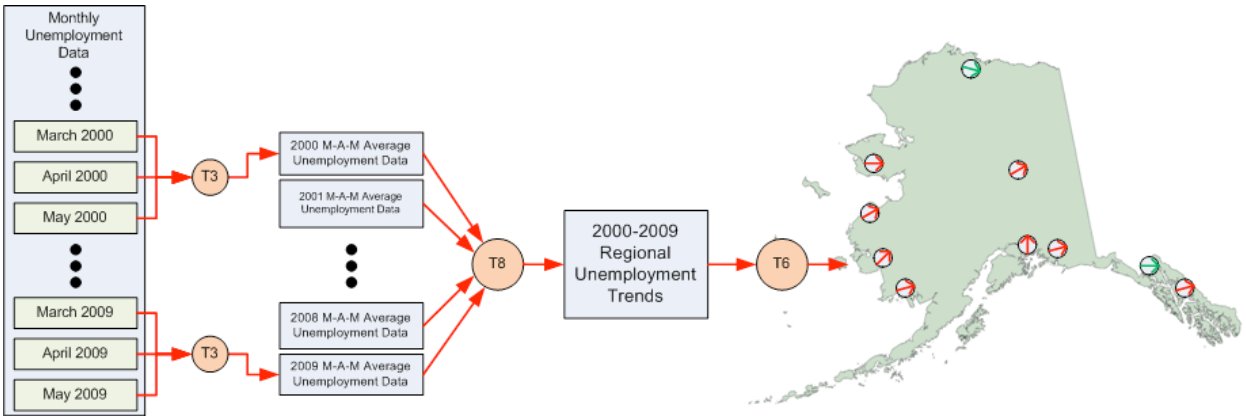

Figure 8 – Reapplication of transformation applied to unemployment datasets to determine trends

5. Conclusions

These two examples clearly demonstrate the need for a more integrated approach to data manipulation and exploration in the scientific community, but as all facets of modern life become increasingly data driven this need extends beyond the scientific realm to almost every aspect of our world. For example, businesses consume data to make more informed decisions about pricing, marketing, production, and demand. Governments use large datasets to determine the needs of the populations they serve, to plan for the future, and to provide national security. Law enforcement agencies are often inundated with digital data during investigations, particularly as computer systems become increasingly involved in not just computer crime, but as source of evidence in more traditional crimes. In many cases the issue is not a lack of data, or even a lack of good data, but rather an inability to effectively and efficiently process, manipulate, explore, and combine the data that does exist to create knowledge.

The PATH Project described in this paper aims to build on the previous work by this project team to ensure that data consumers can easily leverage the work of others to easily build transformation workflows, while also providing the ability for new user-defined transformations to be added to the library. Although still in the prototype phase, initial results are promising and the underlying service-oriented architecture concept which emphasizes reuse to minimize duplication of effort is a promising step in minimizing the efforts associated with merging heterogeneous data sets to facilitate knowledge discovery and contribute to informed decision-making processes.

6. References

[1] Hay, B. and K. Nance. "Simon: An Intelligent Agent For Heterogeneous Data Mapping." Proceedings of the International Conference on Intelligent Systems and Control. Honolulu, Hawaii. August 13-18, 2000.

[2] Nance, K. "Synthesis of Heterogeneous Data Sets." Proceedings of the 9th Annual Software Technology Conference. May 6 – 10, 1997.

[3] Reierson, Lars-Otto. AMAP Secretariat Letter of Commendation, 2001.

[4] Das, S., K. Shuster and C. Wu. "ACQUIRE: agent-based complex query and information retrieval engine." Proceedings of the first international joint conference on Autonomous agents and multiagent systems. Bologna, Italy, 2002.

[5] Di Nitto, Elisabetta , Carlo Ghezzi and Paolo Selvini. "Information access and retrieval: Using agents for multi-target search on the Web." Proceedings of the 2003 ACM symposium on applied computing. March 2003.

[6] Bonifati, A., Chrysanthis, P. K., Ouksel, A. M., and Sattler, K. 2008. Distributed databases and peer-to-peer databases: past and present. SIGMOD Rec. 37, 1 (Mar. 2008), 5-11. DOI= http://doi.acm.org/10.1145/1374780.1374781.