

# Argos: A Framework for Automatically Generating Data Processing Workflows

José Luis Ambite  
University of Southern California, Information Sciences Institute  
4676 Admiralty Way, Marina del Rey, CA 90292  
ambite@isi.edu

Dipsy Kapoor  
University of Southern California, Information Sciences Institute  
4676 Admiralty Way, Marina del Rey, CA 90292  
dipsy@isi.edu

## ABSTRACT

**Demo.** We demonstrate Argos, a framework to *automatically* generate data processing workflows. First, we show how to assign formal semantics to data and operations using to a domain ontology. Specifically, we define data contents using *relational* descriptions in an expressive logic. Second, we show a novel planner that uses relational subsumption to connect the output of a data processing operation with the input of another. Our modeling methodology has the significant advantage that the planner can *automatically* insert adaptor operations wherever necessary to bridge the inputs and outputs of operations in the workflow. We have implemented the approach in a transportation modeling domain.

## Categories and Subject Descriptors

H.2.5 [Information Systems]: Database Management—*Heterogeneous Databases*; H.3.5 [Information Storage and Retrieval]: Online Information Services—*Web-based services*; D.1.6 [Software]: Programming Techniques—*Logic Programming*

## Keywords

Web Service Composition, Workflow, Information Integration, Knowledge Representation, Logic

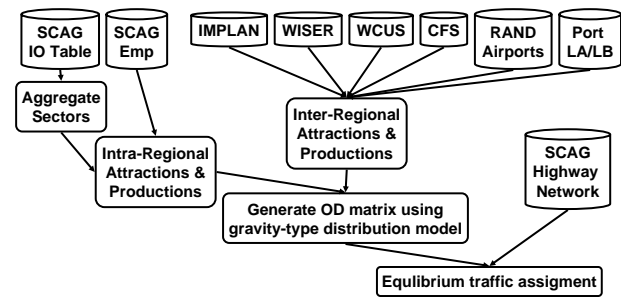
Much of the work of scientists and engineers in government, industry and academia is consumed by accessing, integrating, and analyzing data. Unfortunately, there is a severe lack of tools to facilitate this process. Much of the integration is done manually by ad-hoc methods. Moreover, raw data is of limited utility. Usually raw data are the input to models that produce additional data of interest. For example, in our transportation modeling domain, we derive truck traffic along specific highway links within a metropolitan area, based on far-removed source data such as employment, imports and exports.

The Argos framework improves this state of affairs by (1) describing sources and data processing operations in a way that facilitates sharing and reuse and by (2) generating new data on demand by automatically composing data processing workflows using available sources and operations.

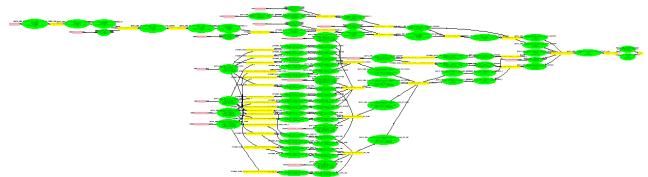
We demonstrate the Argos approach in a transportation modeling domain [3, 2]. However, the Argos methodology is general and can be applied to produce data processing workflows in any other domain, as long as the data and operations are described in a suitable ontology of the domain.

As an example of a data processing workflow consider Figure 1(a) that shows an abstract view of a model [3, 2] for estimating the flow of commodities in a metropolitan

area based on secondary sources. This model was applied to the Los Angeles Consolidated Statistical Metropolitan Area (LACMSA), a region that includes the twin ports of Los Angeles and Long Beach (which together top the U.S. in terms of container shipments), as well as the Los Angeles International airport among others.



(a) Abstract Data Processing Workflow



(b) Structure of the Complete Data Processing Workflow

**Figure 1: Estimating Commodity Flows (LACMSA)**

The model estimates the *intra-regional* trade based on employment data and on an input-output transaction model of the local economy (produced by the Southern California Association of Governments — SCAG), resulting in a table of attractions and productions of different commodity sectors for each traffic analysis zone (TAZ) within the region.<sup>1</sup> To estimate the *inter-regional* trade, the model uses a variety of data sources, including seaborne commerce data from the Waterborne Commerce of the United States (WCUS),<sup>2</sup> airport economic statistics compiled by RAND Corporation,<sup>3</sup> and data from the Commodity Flow Survey (CFS) of the US

<sup>1</sup>A TAZ is a spatial region consisting of several census blocks. The LACMSA is partitioned into 3165 TAZs.

<sup>2</sup><http://www.iwr.usace.army.mil/ndc/data/datawcus.htm>

<sup>3</sup><http://ca.rand.org/stats/economics/airport.html>

Census Bureau,<sup>4</sup> WISER trade,<sup>5</sup> and IMPLAN.<sup>6</sup> The inter-regional trade attractions and productions per commodity are assigned to the TAZs of the entry/exit points in the region. For example, airborne imports of computer equipment are assigned to the TAZs corresponding to the airports in the region. The intra- and inter-regional attractions and productions are converted to an Origin-Destination matrix between pairs of TAZs using a gravity model. Finally, a network equilibrium algorithm assigns the origin-destination data to specific links in the highway network. Figure 2 shows graphically the final result of the workflow: an assignment of the flow of freight to each of the links of the highway network of the region of interest, in our case LACMSA. Thicker lines indicate higher truck traffic volumes.



**Figure 2: Estimated Truck Volume (Los Angeles)**

There is a host of challenges in producing a data processing workflow such as the transportation model described above. Since the data comes from a variety of sources, it may be expressed in different schemas, formats, and units. Therefore, the workflow needs to include many operations that perform different types of data conversion, for example, to translate a given measurement into different units — from tons to dollars to jobs to ton-miles to container units to trucks to passenger-car-equivalents. Also frequent is the need to translate economic data described in one industry/sector classification to another, for example, from the North American Industry Classification System (NAICS) to the Standard Classification of Transported Goods (SCTG), or from different versions of these classifications, for example, from NAICS 1997 to NAICS 2002.

There are many details that the abstract workflow of Figure 1(a) does not show. The workflow that estimates the truck traffic due to freight movements, whose structure appears Figure 1(b), contains over 50 data access and data processing operations. Pan and Gordon implemented this estimation model by a combination of manual steps and custom-designed programs [3]. Argos *automatically* generates such a data processing workflow in response to a user

<sup>4</sup><http://www.census.gov/econ/www/se0700.html>

<sup>5</sup><http://www.wisertrade.org/>

<sup>6</sup><http://www.implan.com/>

data request, including all the necessary data integration and translation operations. A more detailed description of the techniques for automatic workflow generation appears as a research paper in this conference [1].

In our system demonstration, we show the different components of the Argos framework. First, we present an ontology of the application domain, transportation modeling in our case study, and formal descriptions of data sources and operations according to such ontology. Second, we demonstrate our planner automatically generating workflows to answer user data requests. Finally, we display the results of our freight estimation workflow in a GIS tool (cf. Figure 2).

In addition, we demonstrate how workflow operations can be deployed as web services. In Argos, operations can be deployed as web services hosted at different servers in the web, or they can be functions from third-party software libraries. In our system demonstration, we showcase our **Product Conversion** service that automates the translation of data categorized according to different industry/product classifications. We expect that this service will be of significant interest to the economic modeling community. We provide both a web service for programmatic use, as well as a HTML form interface to reach a wider audience.<sup>7</sup>

## Acknowledgments

We thank current members of the Argos group: Genevieve Giuliano, Peter Gordon, Qisheng Pan, LanLan Wang, and JiYoung Park; and alumni: Stefan Decker, Andreas Harth, Mountu Jinwala, Naqeeb Abbasi, Matthew Weathers and Karanbir Jassar; as well as our government partner agencies. Special thanks to Qisheng Pan for contributing his gravity model and freight distribution code to the Argos project.

This material is based upon work supported by the National Science Foundation (NSF) under Award No. EIA-0306905. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of NSF.

## REFERENCES

- [1] Ambite, J. L., and Kapoor, D. Automatic generation of data processing workflows for transportation modeling. In *Proceedings of the 8th Annual International Conference on Digital Government Research (dg.o2007)*, Philadelphia, PA, USA, 2007.
- [2] Giuliano, G., Gordon, P., Pan, Q., Park, J., and Wang, L. Estimating freight flows for metropolitan highway networks using secondary data sources. In *Proceedings of Transportation Research Board Commodity Flow Survey Conference*, Transportation Research Circular, E-C088, pp 154–158, July 2005.
- [3] Gordon, P., Pan, Q. Assembling and processing freight shipment data: developing a gis-based origin-destination matrix for southern california freight flows. Final report of METTRANS research project 99-25, University of Southern California, 2001. [www.mettrans.org/research/final/99-25\\_Final.pdf](http://www.mettrans.org/research/final/99-25_Final.pdf)

<sup>7</sup>Our project homepage <http://www.isi.edu/~argos> links to the HTML interface (<http://altamira.isi.edu:8080/axis/productConversionClient.jsp>) and describes how to use the actual web service (hosted at <http://altamira.isi.edu:8080/axis/ProductConversionService.jws>).