Agile Data Curation Case Studies Leading to the Identification and Development of Data Curation Design Patterns

Karl K Benedict[1], W. Christopher Lenhardt[2], Joshua Wellzie Young[3],
Larissa Chamley Gordon[4], Steve Hughes[5], Suresh Kumar Santhana Vannan[6], Mark Parsons[7]
1) University of New Mexico - kbene@unm.edu, 2) Renaissance Computing Institute, 3) University Corporation for Atmospheric Research, 4) University of Colorado at Boulder, 5) NASA Jet Propulsion Laboratory, 6) Oak Ridge National Laboratory, 7) Rensselaer Polytechnic Institute

IN33C-0136

https://goo.gl/ZURxEJ

Abstract

The planning for and development of efficient workflows for the creation, reuse, sharing, documentation, publication and preservation of research data is a general challenge that research teams of all sizes face. In response to:

- requirements from funding agencies for full-lifecycle data management plans that will result in well documented, preserved, and shared research data products
- increasing requirements from publishers for shared data in conjunction with submitted papers
- interdisciplinary research team's needs for efficient data sharing within projects, and
- increasing reuse of research data for replication and new, unanticipated research, policy development, and public use

alternative strategies to traditional data life cycle approaches must be developed and shared that enable research teams to meet these requirements while meeting the core science objectives of their projects within the available resources.

In support of achieving these goals, the concept of *Agile Data Curation* has been developed in which there have been parallel activities in support of

1) identifying a set of shared values and principles that underlie the objectives of agile data curation,

2) soliciting case studies from the Earth science and other research communities that illustrate aspects of what the contributors consider agile data curation methods and practices, and

3) identifying or developing design patterns that are high-level abstractions from successful data curation practice that are related to common data curation problems for which common solution strategies may be employed.

This paper provides a collection of case studies that have been contributed by the Earth science community, and an initial analysis of those case studies to map them to emerging shared data curation problems and their potential solutions. Following the initial analysis of these problems and potential solutions, existing design patterns from software engineering and related disciplines are identified as a starting point for the development of a catalog of data curation design patterns that may be reused in the design and execution of new data curation processes.

Table 1 - Case Study Information

Id	Purpose	Source	Description
GSToRE	Provide geospatially enabled data discovery, and access services	Benedict, n.d.	A tiered services oriented architecture for providing RESTful and OGC data and metadata services
PDS	The mission of the Planetary Data System (PDS) is to facilitate achievement of NASA's planetary science goals by efficiently collecting, archiving, and making accessible digital data and documentation produced by or relevant to NASA's planetary missions, research programs, and data analysis programs.	Hughes, n.d.	The mission of the Planetary Data System (PDS) is to facilitate achievement of NASA's planetary science goals by efficiently collecting, archiving, and making accessible digital data and documentation produced by or relevant to NASA's planetary missions, research programs, and data analysis programs. The PDS4 Information Model (IM) is a key tool for the management of the PDS4 Information System, a system based on principles from the Open Archive Information System (OAIS). The ingest and use of the BOPPS BIRC Image Data Set is submitted as a case study of Agile Data Curation under PDS4.
WRF	Response sensitivity to increased model resolution on Weather Research and Forecasting (WRF) models.	Gordon, n.d.	the team combined two modeling results, one representing the climate on a century-scale and the other representing high-resolution synoptic dynamics, at two different model resolutions, to study for possible changes in the storm track.
LTER	Harvest and search network node data contents	Porter, n.d. ; Porter, 1997	The DTOC system was extremely simple. Sites posted a DTOC on their web site and told the indexing site where to find that document. Periodically the indexing site would fire off the web crawler to update the search index.
NSIDC	Coevolution of products and users.	Parsons, n.d.; Baker et al. 2016	Over time, NSIDC evolved its suite of sea ice and other products in response to changing user needs and new and emergent user communities.

!

Select References:

Baker, K. S., Duerr, R. E., & Parsons, M. A. (2015). Scientific Knowledge Mobilization: Co-evolution of Data Products and Designated Communities. International Journal of Digital Curation, 10(2). https://doi.org/10.2218/ijdc.v10i2.346

Dubberly, H., & Evenson, S. (2008). On Modeling: The Analysis-systhesis Bridge Model. Interactions, 15(2), 57–61. https://doi.org/10.1145/1340961.1340976

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1995). Design patterns: elements of reusable object-oriented software. Reading, Mass.: Addison-Wesley.

Pence, W. D., Chiappetti, L., Page, C. G., Shaw, R. A., & Stobie, E. (n.d.). Definition of the Flexible Image Transport System (FITS), version 3.0. Astronomy & Astrophysics, 524, A42. https://doi.org/10.1051/0004-6361/201015362

Porter, J. (n.d.). Prototyping Options for All-Site LTER Data Catalog. Retrieved December 11, 2017, from http://www.vcrlter.virginia.edu/dimes/presentations/PORTER/DTOC97/sld001.htm

Russell, N., Ter Hofstede, A. H., Edmond, D., & van der Aalst, W. M. (2005). Workflow data patterns: Identification, representation and tool support. ER, 3716, 353–368.

Zimmerman, J., & Forlizzi, J. (2017). Speed Dating: Providing a Menu of Possible Futures. She Ji: The Journal of Design, Economics, and Innovation, 3(1), 30–50. https://doi.org/10.1016/j.sheji.2017.08.003

Hughes n.d., Gordon n.d., Porter n.d., Benedict n.d., and Parsons n.d. are based on unpublished responses to the Agile Data Curation Case Study online survey

Table 2 - Component Information

Id	Purpose	Source	Description
GSToRE - Data Storage	Management	Benedict, n.d.	File and database storage of data
GSToRE - Metadata Storage	Management	Benedict, n.d.	PostGIS with XML metadata storage
GSToRE - Search	Discovery	Benedict, n.d.	ElasticSearch search indices
GSToRE - Data Access	Access	Benedict, n.d.	OGC (WFS and WCS) and RESTful data access services
GSToRE - Metadata Access	Access	Benedict, n.d.	RESTful metadata access services
GSToRE - Data Visualization	Visualization	Benedict, n.d.	OGC (WMS) and custom RESTful analytic services
PDS - Data Format	Management	Hughes, n.d.	Self-documenting FITS v. 3.0 data format
PDS - User Input	Design	Hughes, n.d.	Integration of user feedback into design and development process
PDS - Workflow	Workflow	Hughes, n.d.	Product generation workflow
WFS - Data Format	Management	Gordon, n.d.	NetCDF data format and generation
WFS - Workflow	Workflow	Gordon, n.d.	Execution of WRF model at multiple resolutions
LTER - Workflow	Workflow	Porter, n.d. ; Porter, 1997	Harvesting of data product links from node web pages published for that purpose
LTER - Search	Discovery	Porter, n.d. ; Porter, 1997	Discovery of network data resources
LTER - Metadata Storage	Management	Porter, n.d. ; Porter, 1997	Storage of harvested metadata
NSIDC - User Input	Design	Parsons, n.d.; Baker et al. 2016	Capture of user needs and integration of those expanding needs into product evolution

!

Join our collaboration by providing your contact information, or a case study through our project's
Open Science Framework Site
https://osf.io/wcxhv/

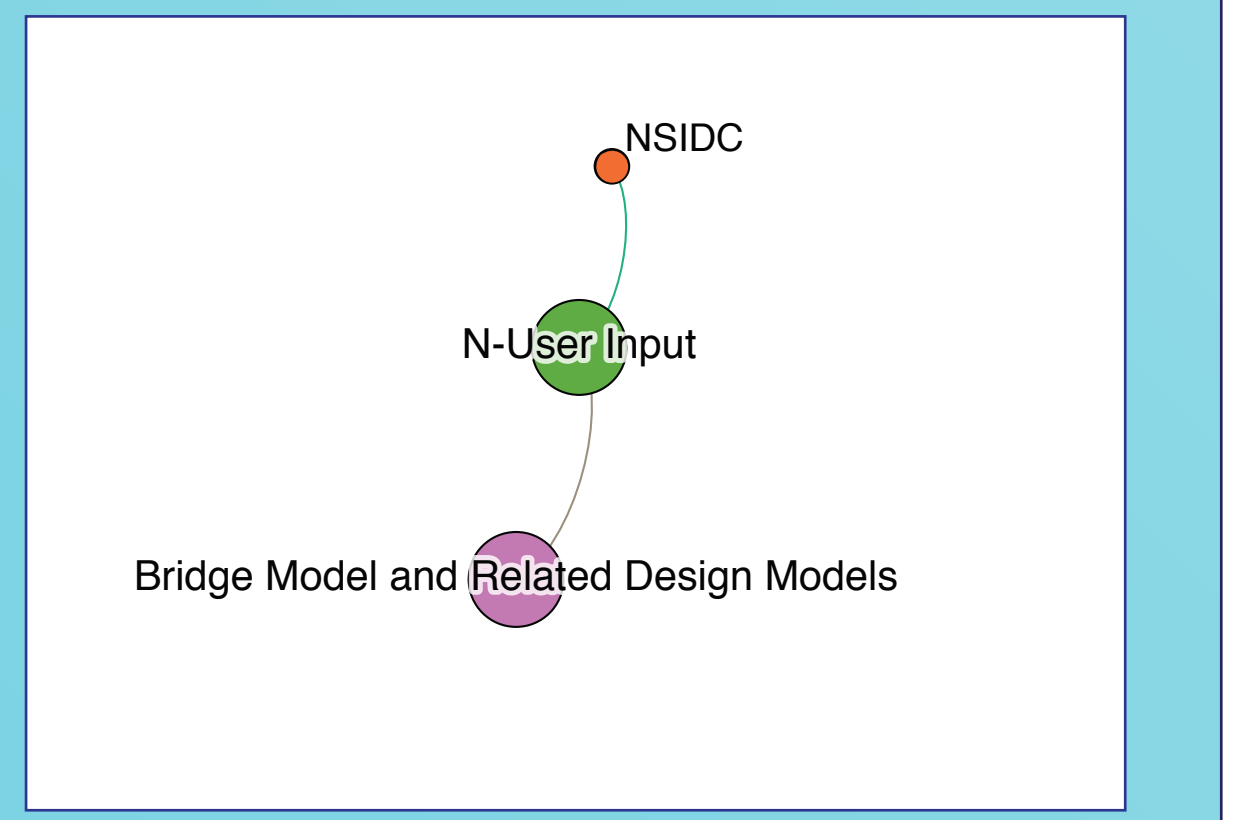

Figure 3 - Free-floating small graph of NSIDC case study

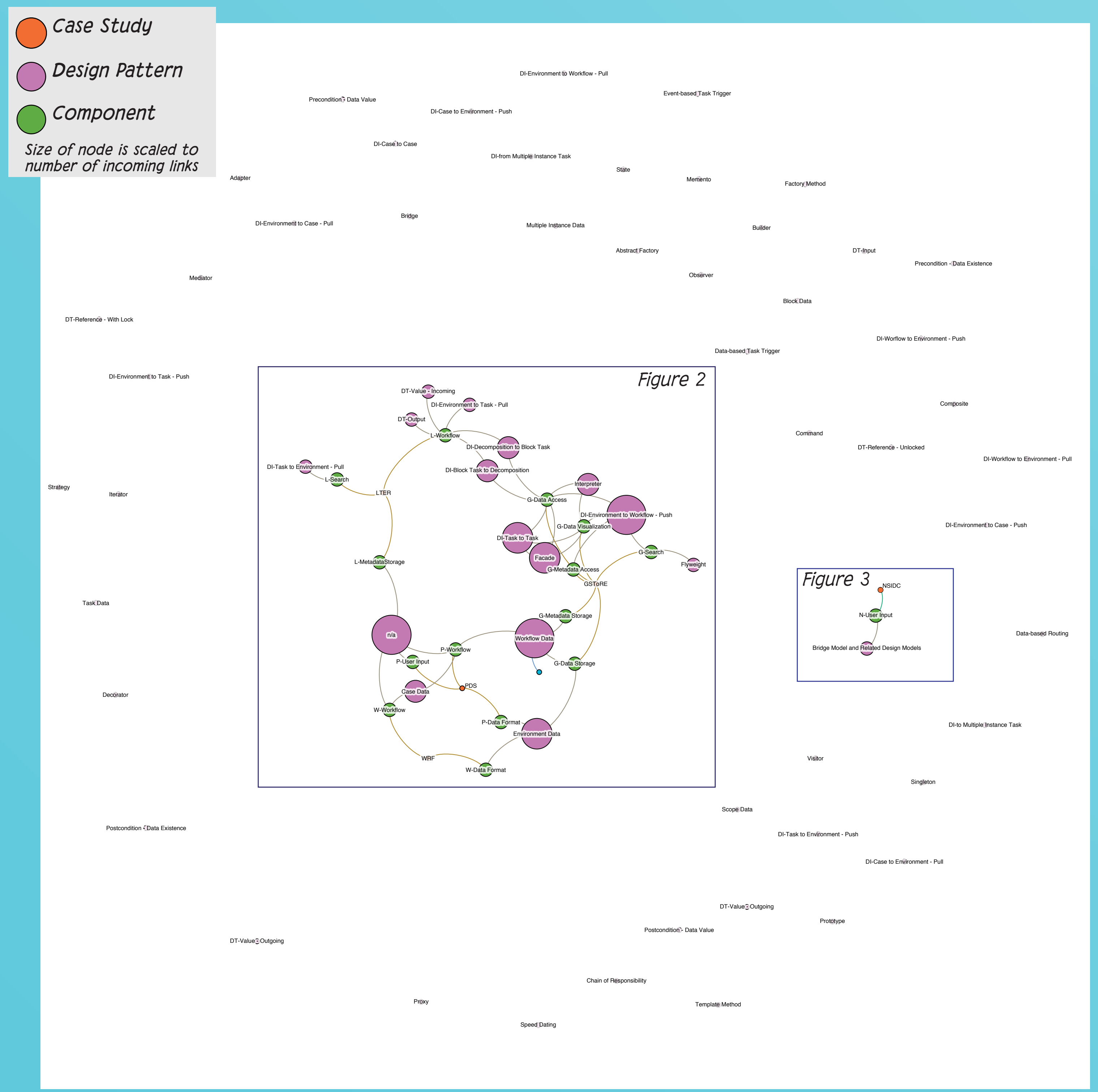

Figure 1 - full graph reflecting both used and unused design patterns, case studies, and associated functional components

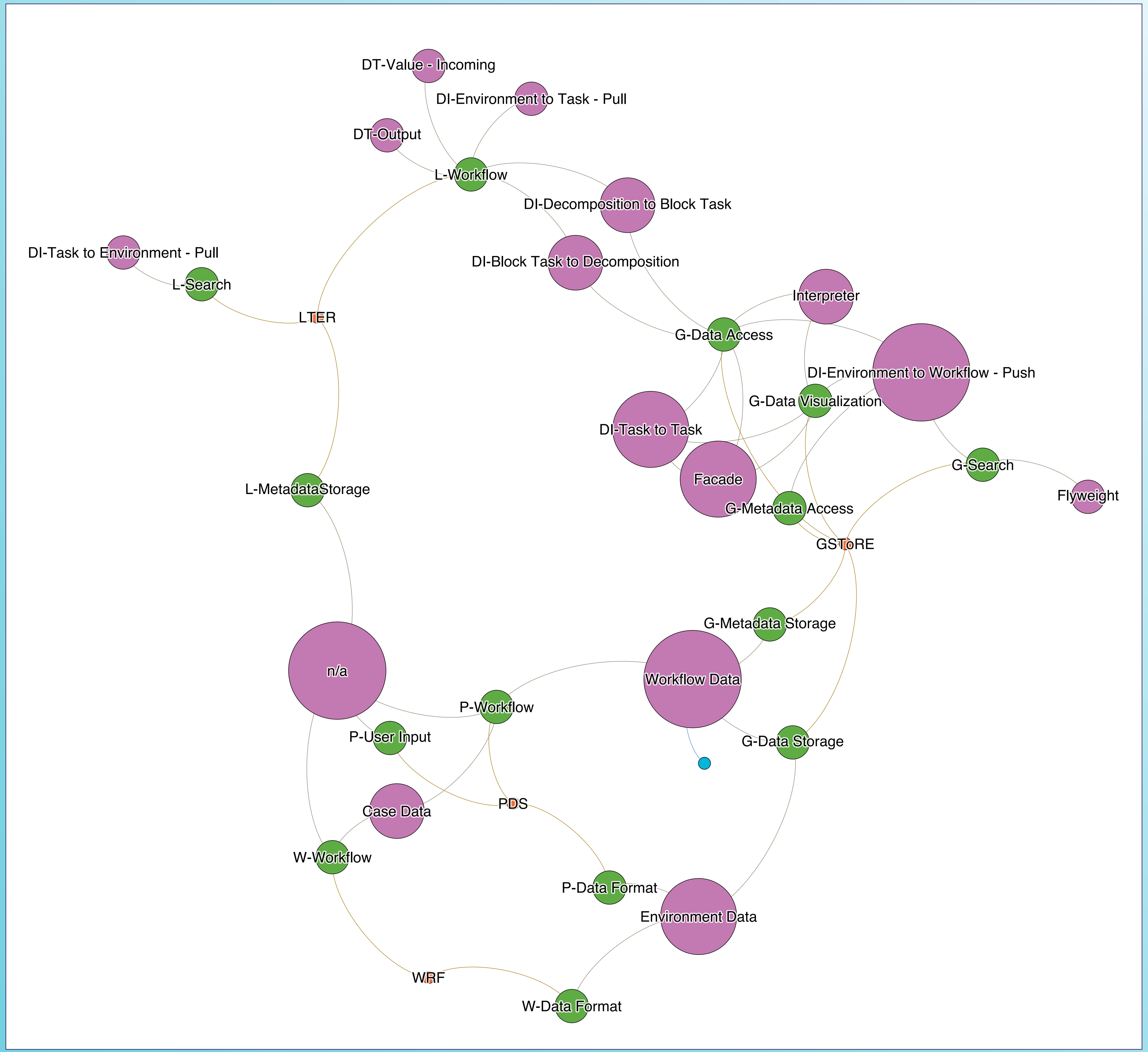

Figure 2 - Graph subset illustrating used design patterns, case studies, and associated functional components

Graph Data Model

The compilation of case studies and linking them to design patterns (either existing or newly developed) is planned as a long-term activity in which examples of effective research data curation practices will be decomponsed into functional components which will then be linked to design patterns. In an effort to facilitate continued expansion of the data model and enable flexible visualiztion, exploration, and analysis of the relationships between case study components and their associated design patterns a *graph data model* was adopted for the project. This model includes *nodes* that represent:

- *Case Studies* contributed by the community
- *Functional Components* associated with those case studies, and
- *Design Patterns* that are linked to those components

Within the data model there are then explicit *edges* defined that represent these connections:

- Case Study -- Functional Component
- Functional Component -- Design Pattern

The overall graph is illustrated in *Figure 1*, with *Figures 2 and 3* providing more detailed perspectives on areas of the graph that illustrate the linkages between the case studies compiled to date and the initial collection of associated components and design patterns.