Community Engagement for Developing the Principles and Practices of Agile Data Curation

Karl Benedict[1] W. Christopher Lenhardt[2] Joshua Young[3]

[1]University of New Mexico [2]Renaissance Computing Institute [3]University Corporation for Atmospheric Research

Abstract

The combination of increasing demands for more systematic data management planning in support of effective in-project data management, research data sharing, and long-term preservation of discovery and access; increasing volumes of data being created and used in research; and research support budgets that aren't increasing in proportion to these demands is creating a situation in which greater efficiencies and product-centered research data management workflows and processes are needed. Taking inspiration from the principles and practices of agile software development, the presenters of this poster are working towards the development of three sets of interdependent products. First, a set of *core principles* that have broad support within the community will be identified and/or developed from existing statements of principles; solicited from the broad community of research data creators, curators, and users; and derived from implicit principles exemplified by specific research data projects that have achieved notable success in enabling efficient use, preservation, discovery and reuse. Second, building upon the case studies identified and reviewed as part of the aforementioned process of identifying core data management principles, additional case studies are being sought that demonstrate effective research data management practices that are aligned with the identified principles and have resulted in well structured, effectively preserved, and documented data sets that are well-positioned for discovery and reuse by diverse users. And third, the development of a collection of research data curation design patterns that can provide structured guidance to researchers and data curators in developing workflows that deliver incremental increases in research data value through time, both to the researchers who are creating and using data for the first time and for future users of those data. This progression from values and principles through current exemplars to documented recommended practices *that are of sufficient specificity to be actionable* is anticipated to produce the theoretical *and* operational foundation needed for more efficient development and delivery of research data value. This greater efficiency has the potential to allow for the continued growth of the volume, diversity, and rate of creation within limited resources, while also enabling more effective collaboration among increasingly large and diverse research teams.

Introduction

Thus far the focus of the project's work has been on developing a framework within which the team can discuss the concept of *agile data curation* with the community, and iteratively evolving that framework through a series of meeting sections, workshops and presentations that have been given at multiple venues including the American Geophysical Union (2014, 2015), Federation of Earth Science Information Partners Meeting (2016), Research Data Alliance (2014, 2015, 2016), and SciDataCon (2016). In these various activities the team has worked on communicating the conceptual framework for our vision of agile data curation, presented a variety of initial values and principles derived from those defined in the *Manifesto for Agile Software Development* (Beck et al., 2001), and solicited the presentation of data management projects that exemplify (either intentionally or unintentionally) these principles. The purpose of the presentation below is to expand this discussion to a broader international and disciplinary audience in support of moving forward with soliciting input into the definition of a set of shared values and principles, and collection of illustrative case studies in support of the development of design patterns that may be applied to diverse data curation problems.

Values and Principles

At the foundation of a conceptual mapping between *Agile Software Development* and *Agile Data Curation* the primary focus is not on the various agile software development methodologies that have been developed, but instead on the underlying values and principles (Beck et al., 2001) that have been identified as a foundation for multiple methodologies identified as *agile*. Below are some initial *agile data curation* values and principles that the authors have developed as a point of departure for a community discussion.

Mapping of Agile Software Development Values into Data Curation

Agile Software Development

- *Individuals and interactions* over processes and tools
- *Working software* over comprehensive documentation
- *Customer collaboration* over contract negotiation
- *Responding to change* over following a plan[^agilePrinciples]

Agile Data Curation

- *Individuals and interactions* over processes and tools
- *Discoverable, understandable and usable data* over comprehensive documentation
- *User collaboration* over contract negotiation
- *Responding to change* over following a plan

Mapping of Agile Software Development Principles into Data Curation

Agile Software Development Principles

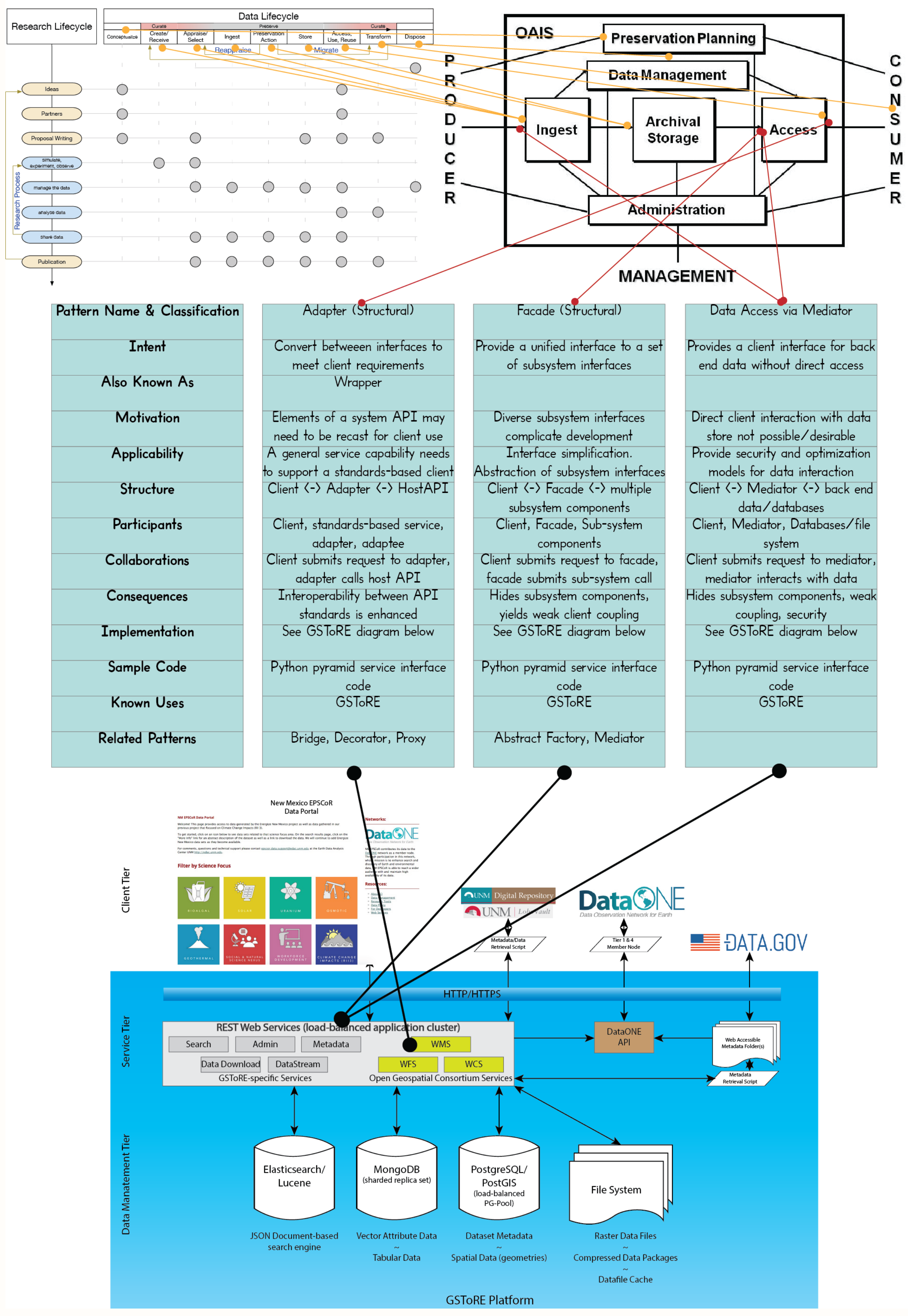
- Our highest priority is to satisfy the customer through early and continuous delivery of valuable software.
- Welcome changing requirements, even late in development. Agile processes harness change for the customer's competitive advantage.
- Deliver working software frequently, from a couple of weeks to a couple of months, with a preference to the shorter timescale.
- Business people and developers must work together daily throughout the project.
- Build projects around motivated individuals. Give them the environment and support they need, and trust them to get the job done.
- The most efficient and effective method of conveying information to and within a development team is face-to-face conversation.
- Working software is the primary measure of progress.
- Agile processes promote sustainable development. The sponsors, developers, and users should be able to maintain a constant pace indefinitely.
- Continuous attention to technical excellence and good design enhances agility.
- Simplicity–the art of maximizing the amount of work not done–is essential.
- The best architectures, requirements, and designs emerge from self-organizing teams. At regular intervals, the team reflects on how to become more effective, then tunes and adjusts its behavior accordingly.

Agile Data Curation Principles

- Maximize the impact of research data through accelerated capacity for discovery, access and use of valuable data
- Expect unanticipated needs for and uses of research data (and documentation) and develop flexible systems to support new uses and users without significant modifications
- Facilitate automated interaction with data and metadata assets through well documented public web services that enable disintermediated use and reuse of research data
- Data creators and data curators should work closely throughout planning, research and preservation activities to ensure the most efficient and streamlined process
- Identify key individuals in a data curation project that have the requisite knowledge and motivation to do the job and get out of their way
- Identify the most effective method(s) for maintaining close communication and *use* them
- Delivery, access, use and citation of research data are the primary measures of success
- Design principles that enable steady delivery of incremental improvements to research data discovery, access and use should be consistent with a sustainable level of effort and funding from sponsors, data creators and curators, and users
- Continuous attention to technical excellence and good design enhances agility
- Start with the basics and only make systems more complex as needed, while maintaining a low bar to entry
- Continuously work to develop and evolve a community of data providers, curators and users that all participate in the ongoing evolution of the research data systems that they interact with

Case Studies into Design Patterns

Below is an illustration of an initial set of existing design patterns from a software development context that can provide elements for addressing elements of a combination of research and data lifecycles, the OAIS archival framework, and the capabilities of a specific data management, discovery and access platform (Benedict, 2017); the Geographic Storage, Transformation and Retrieval Engine - GSToRE - (bottom - Earth Data Analysis Center (EDAC), 2016) developed in support of a specific set of research and application requirements



Pattern Name & Classification	Adapter (Structural)	Facade (Structural)	Data Access via Mediator
Intent	Convert betweeen interfaces to meet client requirements	Provide a unified interface to a set of subsystem interfaces	Provides a client interface for back end data without direct access
Also Known As	Wrapper		
Motivation	Elements of a system API may need to be recast for client use	Diverse subsystem interfaces complicate development	Direct client interaction with data store not possible/desirable
Applicability	A general service capability needs to support a standards-based client	Interface simplification. Abstraction of subsystem interfaces	Provide security and optimization models for data interaction
Structure	Client <-> Adapter <-> HostAPI	Client <-> Facade <-> multiple subsystem components	Client <-> Mediator <-> back end data/databases
Participants	Client, standards-based service, adapter, adaptee	Client, Facade, Sub-system components	Client, Mediator, Databases/file system
Collaborations	Client submits request to adapter, adapter calls host API	Client submits request to facade, facade submits sub-system call	Client submits request to mediator, mediator interacts with data
Consequences	Interoperability between API standards is enhanced	Hides subsystem components, yields weak client coupling	Hides subsystem components, weak coupling, security
Implementation	See GSToRE diagram below	See GSToRE diagram below	See GSToRE diagram below
Sample Code	Python pyramid service interface code	Python pyramid service interface code	Python pyramid service interface code
Known Uses	GSToRE	GSToRE	GSToRE
Related Patterns	Bridge, Decorator, Proxy	Abstract Factory, Mediator	

Figure 1: Mapping of the GSToRE Platform's Capabilities into a Set of Design Patterns (middle - Gamma et al., 1995, pp 135, 185; Schwinn and Schelp, 2005, pp 476) and corresponding linkages between the OAIS Framework (upper right - Consultative Committee for Space Data Systems (CCSDS), 2012; International Organization for Standardization (ISO), 2012; Lavoie, 2014) and an Illustration (upper left) of the Intersection of the Research Lifecycle (JISC, 2014) and Data Curation Lifecycle Actions (Digital Curation Centre (DCC), nd)

Bibliography

Beck, K., Beedle, M., van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., Grenning, J., Highsmith, J., Hunt, A., Jeffries, R., Kern, J., Marick, B., Martin, R.C., Mellor, S., Schwaber, K., Sutherland, J., Thomas, D., 2001. Manifesto for Agile Software Development.

Benedict, K., 2017. The Geographic Storage, Transformation and Retrieval Engine (GSToRE): A Platform for Active Data Access and Publication as a Complement to Dedicated Long-Term Preservation System, in: Curating Research Data. Volume Two, A Handbook of Current Practice. Association of College and Research Libraries, Chicago, IL, pp. 207–209.

Consultative Committee for Space Data Systems (CCSDS), 2012. Reference Model for an Open Archival Information System (OAIS) (No. CCSDS 650.0-M-2). Consultative Committee for Space Data Systems (CCSDS).

Digital Curation Centre (DCC), nd. DCC Curation Lifecycle Model | Digital Curation Centre.

Earth Data Analysis Center (EDAC), 2016. GStore V3 API.

Gamma, E., Helm, R., Johnson, R., Vlissides, J., 1995. Design patterns: Elements of reusable object-oriented software, Addison-wesley professional computing series; addison-wesley professional computing series. Addison-Wesley, Reading, Mass.

International Organization for Standardization (ISO), 2012. ISO 14721:2012 - Space data and information transfer systems – Open archival information system (OAIS) – Reference model. ISO.

JISC, 2014. How Jisc is helping researchers : Jisc.

Lavoie, B., 2014. The Open Archival Information System (OAIS) Reference Model: Introductory Guide (2nd Edition). Digital Preservation Coalition.

Schwinn, A., Schelp, J., 2005. Design patterns for data integration. Journal of Enterprise Information Management 18, 471–482. doi:10.1108/17410390510609617

Acknowledgements

This work has been partially supported through funding from the National Science Foundation (#IIA-1301346)

International Digital Curation Conference. Edinburgh, Scotland. February 20-23, 2017.