# Improving Traceability Throughout the Data Lifecycle: the DOLCE Approach to Provenance

Lydia A. M. Fletcher
Texas Advanced Computing Center
University of Texas at Austin

Paul Buschow
Center for Space Research
University of Texas at Austin

Greg Smithhart
Center for Space Research
University of Texas at Austin

## Abstract

The Texas Advanced Computing Center (TACC) uses a while lifecycle approach to data management called the Digital Object LifeCycle Ecosystem (DOLCE). We use DOLCE as a framework to create policies and services that enable TACC to support accessibility to and discovery of digital objects throughout the phases of their lifecycle including generation, processing, description, analysis, storage, and sharing. The ultimate goal of DOLCE is to produce data that aligns with the FAIR Principles of findability, accessibility, interoperability, and reusability. A key aspect of this data curation process is using data provenance to improve reusability of data by tracking the processes used to create, gather, transform, and analyze digital objects. In this poster, we present our goals for improving data provenance using geospatial data as a use case. We will demonstrate how we utilize robust metadata to capture important data processing steps. We will also explain how provenance ties into our development of a data catalog that promotes long-term preservation and reusability.

## Overview of DOLCE

The Data Management and Collections (DMC) team at the Texas Advanced Computing Center (TACC) supports data collection, curation, and preservation processes to support research data management (RDM) using the Digital Object Life-Cycle Ecosystem (DOLCE) framework. The overall goal of DOLCE is to create policies and services that enable TACC to support accessibility to and discovery of digital objects throughout the phases of their lifecycle . These phases include:

- *Generation* - The creation of new data through direct observation, development of new models, or other primary research activities.
- *Processing* - Cleanup or other refining of data products including activities such as data transformation, removal of personally identifiable information, encryption, or compression.
- *Description* - Creation of robust metadata to facilitate long term discoverability, access, and reuse.
- *Analysis* - Use in scientific analysis to inform interpretation and gain meaningful insights from the digital objects. This includes using datasets for statistical analysis, as input for machine learning algorithms, to develop visualizations, and as input for modeling activities.
- *Storage* - Both short term storage for immediate access and long-term preservation.
- *Sharing* - Making the data available beyond the lifespan of its current project for reuse by other researchers in the short term or in the future.


*Figure 1: Sample overview of a digital object's lifecycle*

## Overview of Provenance

Data provenance has emerged as a critical concept and practice that addresses concerns about data quality, integrity, and reliability. Data provenance refers to the documented history and origin of data, tracing its lifecycle from creation to final use [Buneman 2001]. It encompasses information about the processes, entities, and activities involved in the data's journey, enabling comprehensive understanding and analysis of its reliability, trustworthiness, and quality. Provenance captures metadata that reveals why, how, where, when, and by whom data was created, modified, accessed, and transformed, creating an audit trail for data assets [Magagna 2020].

In scientific research and experimentation, data provenance plays a fundamental role in facilitating reproducibility. By capturing the entire history of data processing, analysis, and results, researchers can replicate experiments precisely, validate findings, and foster collaboration [Gil 2016]. Provenance also aids in error detection, allowing researchers to identify and rectify discrepancies, contributing to the overall advancement and transparency of scientific inquiry [Bao 2021].

As data sources proliferate, provenance becomes instrumental in achieving effective data integration and interoperability. By capturing data lineage and transformations, provenance enables data consumers to understand data semantics, context, and dependencies, facilitating seamless integration across diverse systems and improving data interoperability either by aligning metadata collection with FAIR principles or with other international standards [Magagna 2020]. In keeping with the FAIR principles, efforts should be made to ensure that captured provenance information is machine-readable [Closa 2017].

Establishing the context of a given digital object is also key to the successful assessment and re-use of that digital object by other researchers, decision-makers, or the general public. By providing visibility into data lineage, source authenticity, and data transformation operations, data provenance serves as a vital tool for assessing data quality and ensuring its reliability by establishing the contextual information that enables assessment [Faniel 2019]. While the FAIR principles outline best practices for data creators and providers, many data re-users rely on "shortcuts" such as the level of trust they have in the reputation of a data creator or provider [Faniel 2019; Bishop and Collier 2022], which well-documented provenance enables.

Provenance information instills confidence in data-driven decision-making processes, promoting trust among stakeholders and minimizing the risk of erroneous or misleading conclusions [Bao 2021]. By providing a transparent view of data origins, transformations, and decision-making processes, provenance enables stakeholders to evaluate the validity and fairness of decisions, promoting trust and ethical practices.

DOLCE's approach to provenance is to create services and workflows that encourage the collection of provenance information at each stage of a digital object's lifecycle, while maintaining awareness of the needs of data reusers.

## Current Efforts to Capture Provenance

Members of the TACC DMC team and partners with the UT Center for Space Research (CSR) engaged in exploring capturing provenance information as part of our work on the Texas Disaster Information System (TDIS). TDIS was designed as an interactive, web-based spatial data system designed to support disaster preparedness, response, recovery, and mitigation for the State of Texas. A key component of TDIS involves stewarding high-value disaster datasets and employing data analytics and mapping technologies to provide the information needed to understand hazard risks, assess disaster impacts and develop mitigation strategies. TDIS provides a rational system for capturing, cataloging, and maintaining information by utilizing TACC's DOLCE framework.

A TDIS-specific metadata standard was developed to create descriptive and other forms of metadata to enable discoverability, access, interoperable use, and dataset reuse. The standard that sets out the metadata fields we expect to collect for datasets, models, and other artifacts within collections and is intended to be flexible enough to cover a wide variety of digital objects that we expect will ingest into TDIS. This information is published via GitHub: https://github.com/TexasDIS/metadata

In developing the TDIS metadata schema, we interpreted provenance to encompass what we have defined as processing steps in addition to any important activities related to the management and stewardship of a resource by one or more parties over time. We developed fields in the schema to describe any changes made by successive custodians of a given digital object. This aligns the TDIS metadata schema with the Federal Geographic Data Committee standard FGDC-STD-001-1998.

Within the TDIS schema we categorized fields according to their primary purpose – for example, the grouping "contact information" covers the fields related to the primary point of contact for a given digital object. The "reuse" grouping is intended to capture not only information about prior reuse of a digital object, but also aid in making decisions regarding whether a digital object is appropriate for reuse in a new project. It includes repeatable fields such as "chain of custody" to identify any "changes in ownership and custody of the digital object since its creation that are significant for its authenticity, integrity, and interpretation." The reuse grouping also includes various repeatable fields related to any processing steps that may have been undertaken that transforms the structure of the data.

**Reuse**

**Field Name:** Chain of Custody
**Description:** A statement of any changes in ownership and custody of the digital object since its creation that are significant for its authenticity, integrity, and interpretation. The statement may include a description of any changes successive custodians made to the digital object.
**Use:** Recommended
**Conditional:** No
**Use Condition:** None
**Accepts Multiple Values:** Yes
**Format:** Text
**Controlled Terms URL:** None

*Figure 2: Example metadata field from TDIS metadata schema*

## Future Efforts

A key topic of current discussion regarding provenance is how to convey the trustworthiness of a digital object. Many users and reusers of digital objects rely on their knowledge of a data producer's or organization's reputation to make decisions regarding a digital object. However, we anticipate working with a variety of end users who may not have the same level of familiarity with key data producing organizations in the state of Texas.

Therefore, we foresee the need to quantify trustworthiness by establishing metrics for assessment – for example, to establish how many digital objects an individual or organization has contributed and the completeness or quality of their metadata. One method for accomplishing this is to establish a set of queries against the TDIS data catalog and pair this with a scoring method that will enable us to signal within the UI our assessment of a digital object's trustworthiness.

The TACC DMC and Decision Support Office teams also work closely with the Model Integration (MINT) team to develop a holistic model and data system and we anticipate continuing discussions of using their WINGS and PEGASUS workflows and tools to enhance the decision support system we are building.

## Acknowledgements

Early development of the TDIS metadata schema was principally undertaken by Anna J. Dabrowski, formerly of TACC, and Brent Porter of UT's Center for Space Research.

Scan for full text and bibliography.

TEXAS

TACC
TEXAS ADVANCED COMPUTING CENTER