

National Oceanic and
Atmospheric Administration

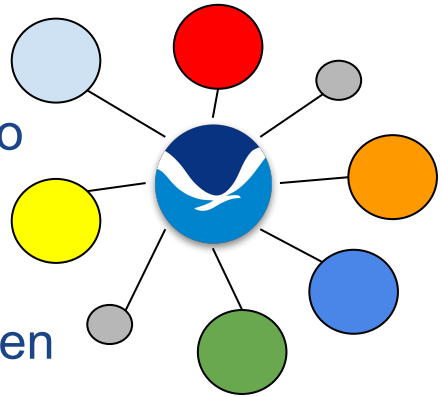
Open Data Curation in the Cloud: Facilitating Transparency through Transformation of NCEI's Data Stewardship and Curation Practices

105th Annual AMS
January 13, 2025
New Orleans, LA

Nancy A. Ritchey, Kenneth S. Casey, PhD
NOAA National Centers for Environmental Information

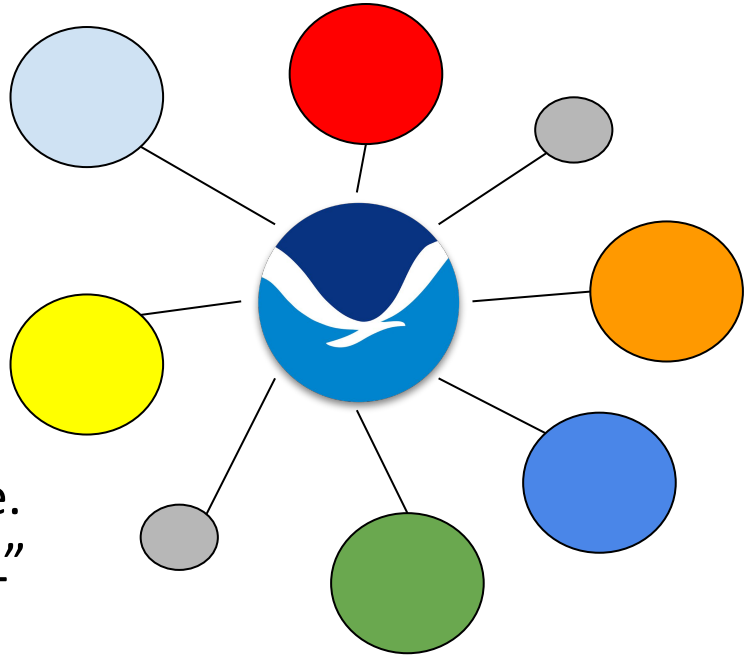
"... facilitating transparency of science and leveraging community supported open science capabilities to drive research and outcomes …"

- NOAA/NCEI is building the **Open Information Stewardship Service (OISS)**, our new archive in the cloud
- Learned the cloud environment is critical in allowing us to operate at the speed and scale needed to meet today's demands
- Utilizing technologies of the semantic web and linked open data to enable open science, AI-based applications, etc.
- This new approach will also enable open curation



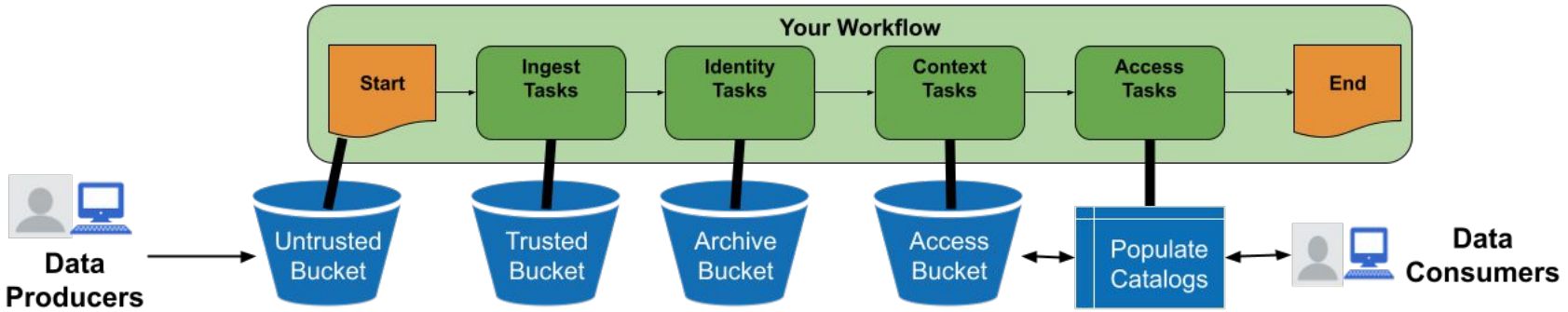
So, What is OISS?

The *simplest* answer is that OISS is the data workflow manager that we are using to archive and steward data in the cloud, and that our science teams can use, if they choose, to *fully* support their science workflows too.



OISS = Open Information Stewardship Service.
Pronounced “oh-is”, like “Lois” without the “L”

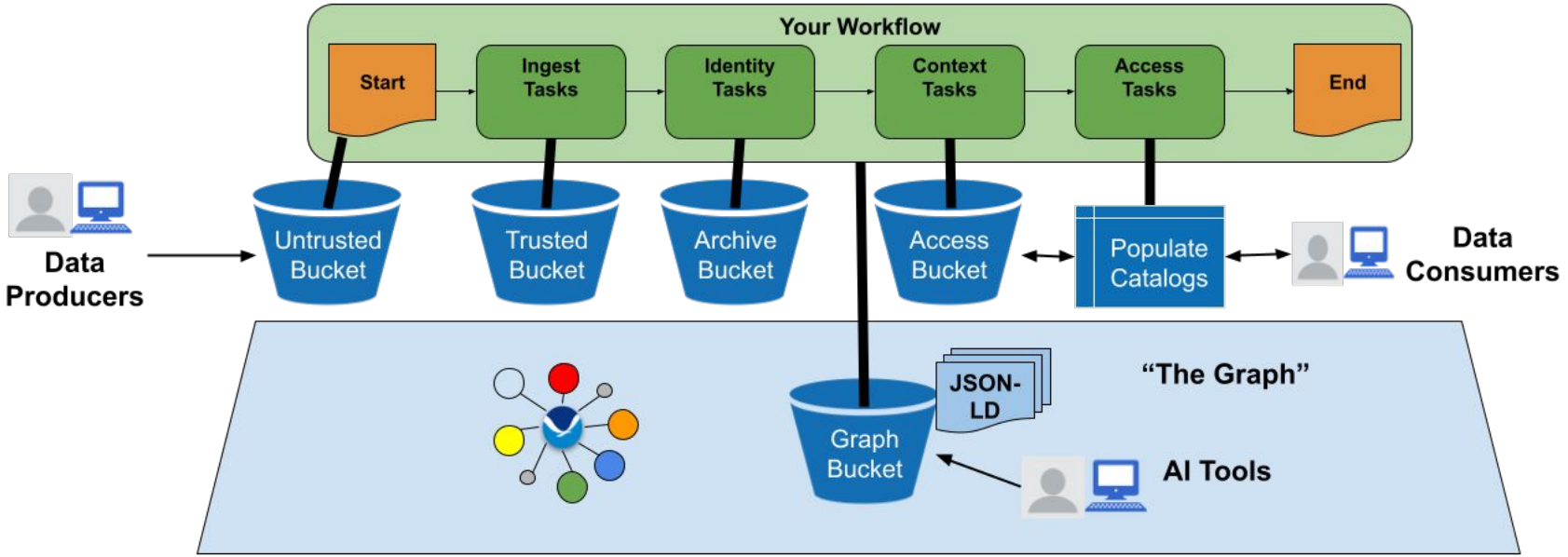
OISS as a Workflow Manager



Pretty straightforward…. and the cloud itself is very helpful here, allowing us do this for many concurrent data streams with very low latency

OISS and Knowledge Graph (KG)

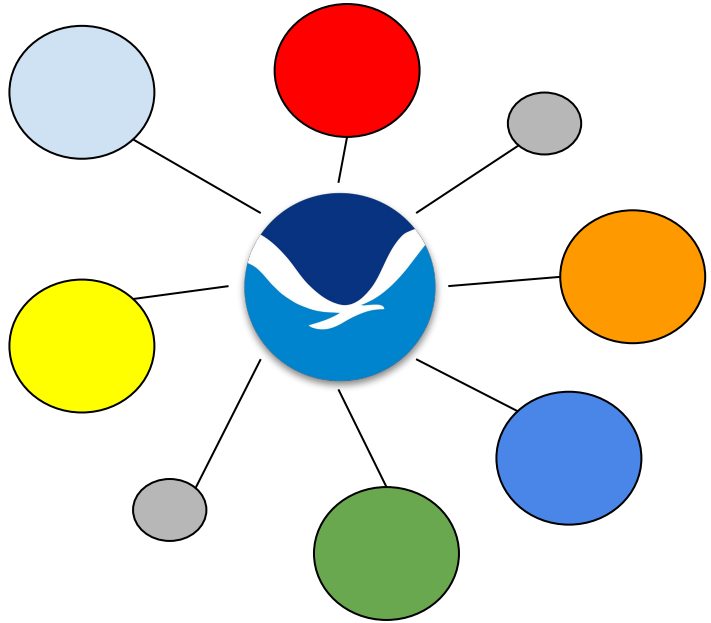
Add “The Graph” underneath, and you have something special…



The workflow, and the result of every step, is logged in the graph

Why is the KG so special?

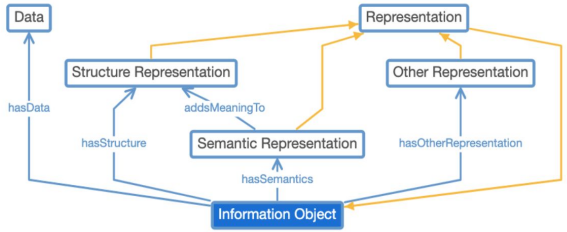
- Each workflow step writes its results to the graph, and the actual workflow is stored in the graph, you get full provenance and transparency… that is, ***Scientific Reproducibility****
- As important as writing the results and the workflow to the graph is *how* they are written, as fully self-documenting JSON-LD documents that machines and people alike can harvest… that is, ***AI-enablement, linked open data, and more…***



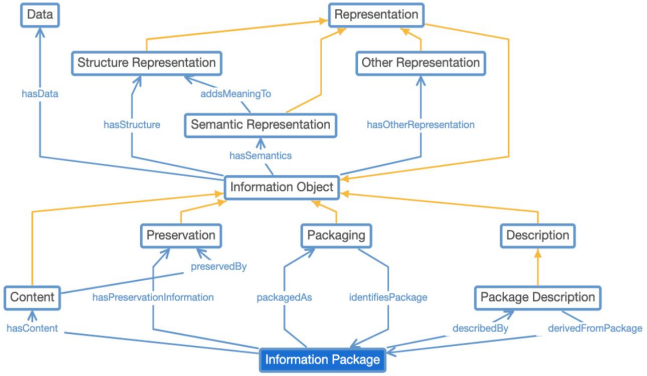
* now required by US federal policy

The “Special Sauce”: Semantic Interoperability

- “**Process as data**” is big step forward, but alone cannot solve interoperability
- Semantic interoperability within a shared reference model is required to allow data and process information to be shared and reused by people and their applications across communities and domains
- Using terms and patterns from existing ontologies (SSN, SOSO, Schema.org, SensorML, Dublin Core, GCMD, NIEM, SWEET, etc.) enables users of the cloud archive service to reap rich benefits of the semantic web
- Enable new, user-defined ontologies constructed and published automatically to the KG for discovery and reuse

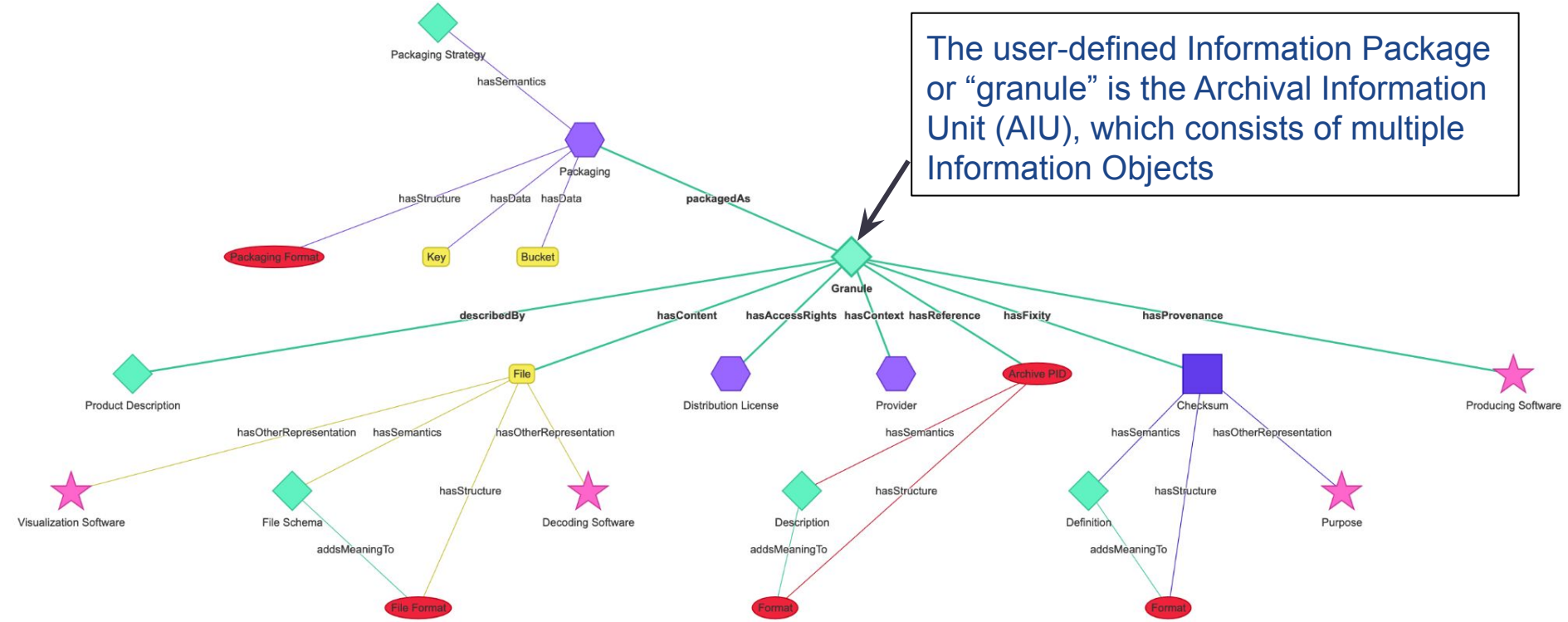


Information Object: primary building block, recursive and generic



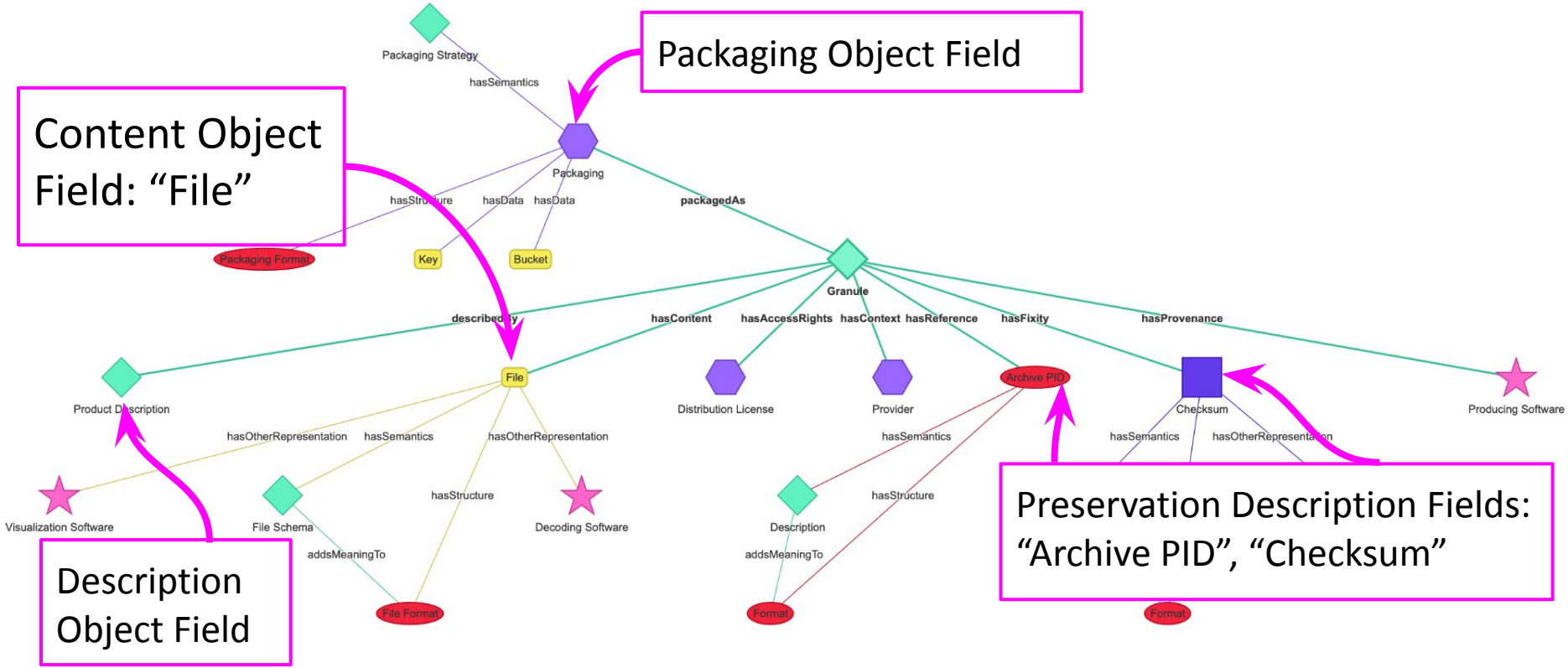
Information Package: secondary building block, a complete generic model

Real World Example: Visualization of the JSON-LD



The user-defined Information Package or “granule” is the Archival Information Unit (AIU), which consists of multiple Information Objects

Real World Example: Visualization of the JSON-LD



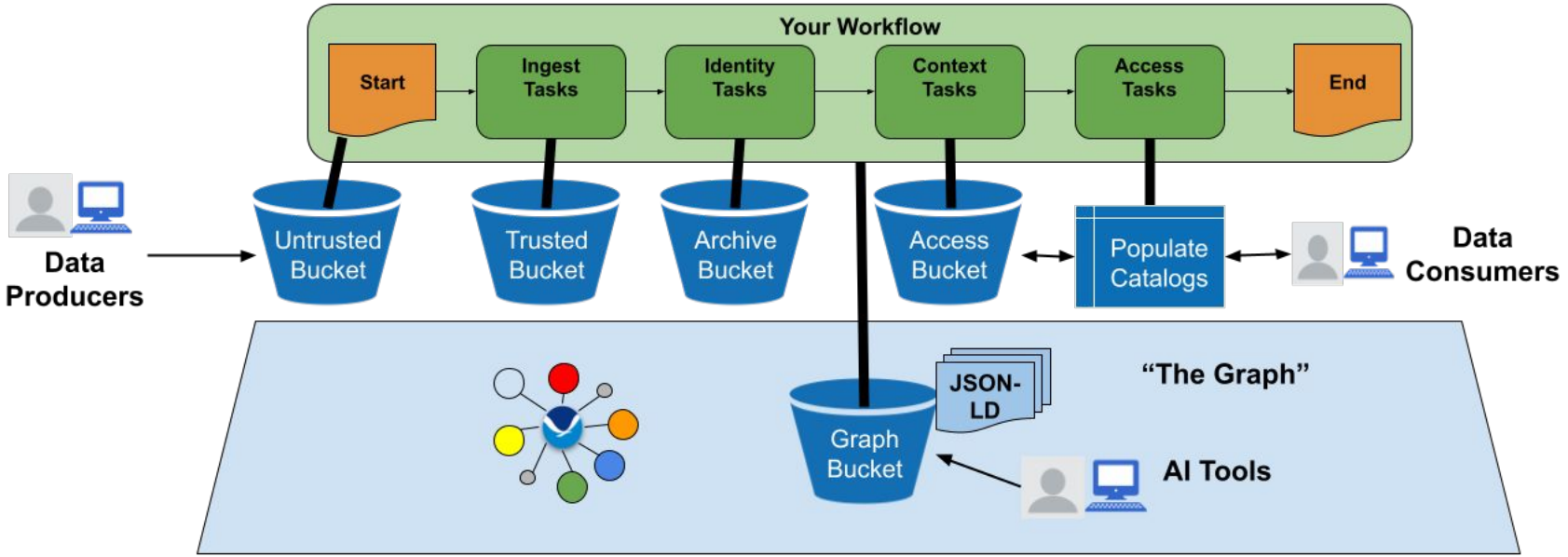
How does this enable open curation?

- Custom fields can be added to the graph as needed, and the structure of these objects can/will evolve over time
  - Additional data, services, feedback, …

- New knowledge, not explicitly entered into the graph, can be inferred through the relationships documented in the graph
  - Related data, publications, organizations, …

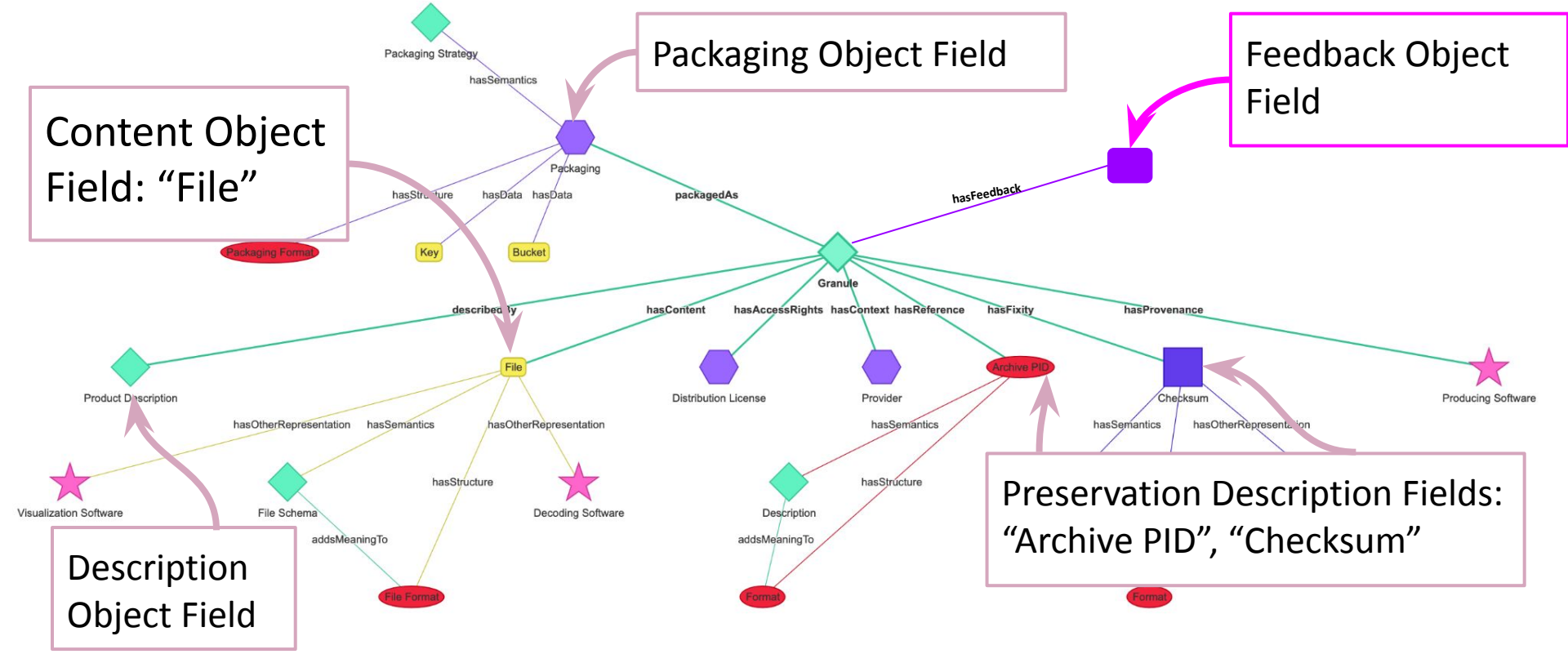
Potential Feedback Option

Workflow for feedback from users can be added to the KG…



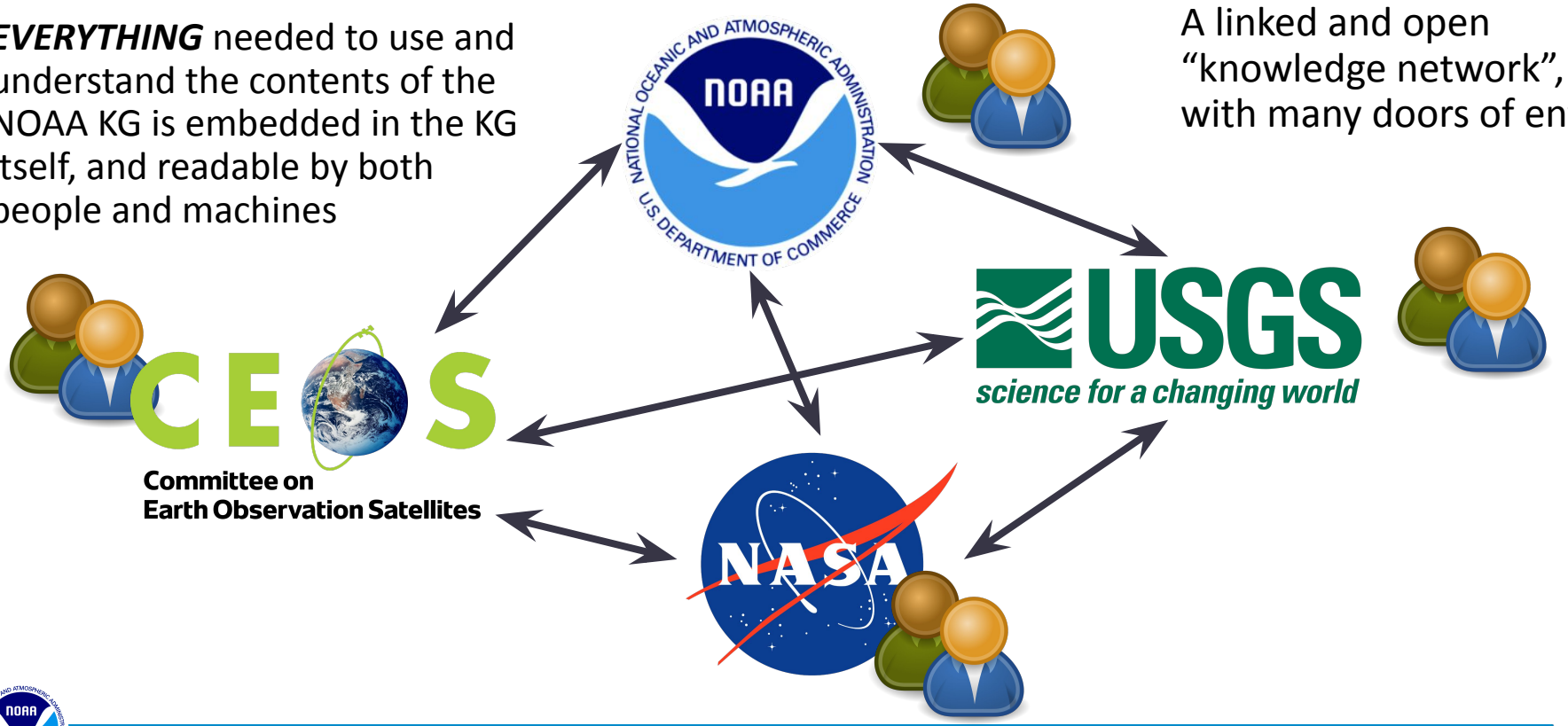
logging the new information in the graph

Real World Example: Visualization of the JSON-LD



Imagine a world of interconnected Knowledge Graphs!?

EVERYTHING needed to use and understand the contents of the NOAA KG is embedded in the KG itself, and readable by both people and machines



A linked and open "knowledge network", with many doors of entry



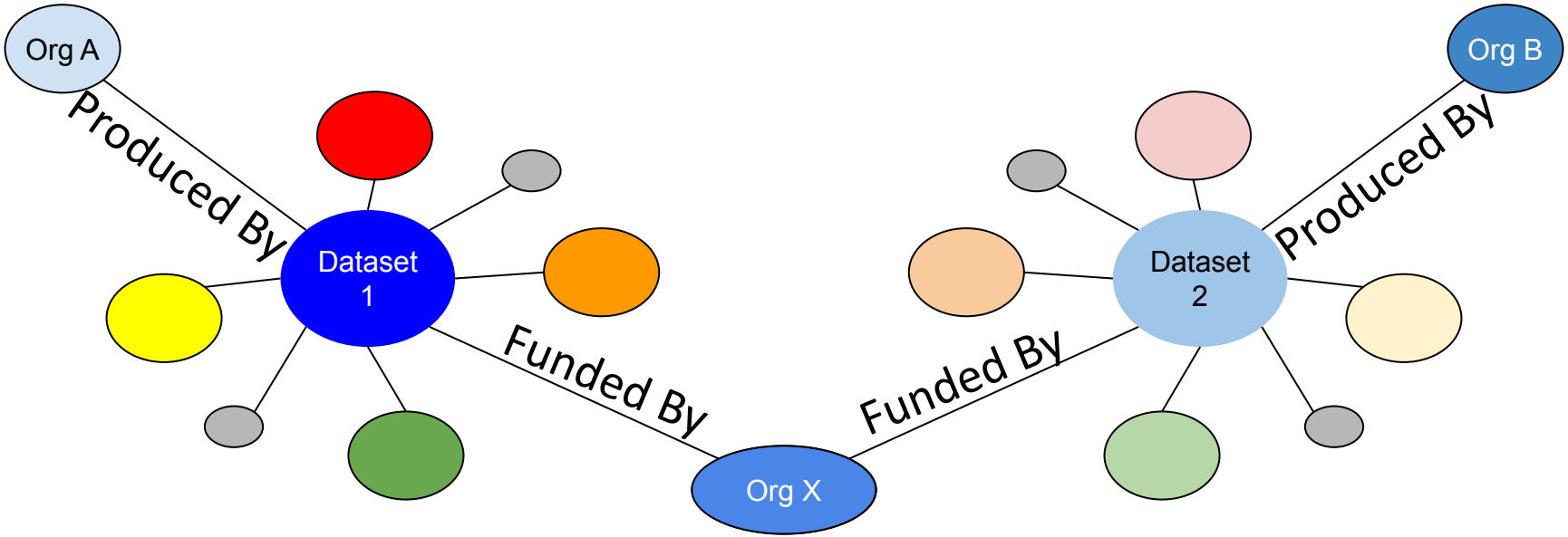
Everything, Everywhere, All at once, Accessible By Everyone, leading to Unanticipated Discoveries!

Questions



©GettyImages-Rowan Jordan

Greater than the sum of its parts



New knowledge, not explicitly entered into the graph, can be inferred through the relationships documented in the graph…. for example, “Org A and Org B are funded by Org X” can be gleaned by “traversing the graph”.