# Development and utilization of data curation process ontology

Inter-University Research Institute Corporation / Research Organization of Information and Systems
National Institute of Informatics
https://www.nii.ac.jp/en/

## Overview

We introduce the data curation process ontology for sharing data curation tasks and procedures across fields. Our proposed ontology provides a well-formalized structure for the data curation activity and may function as a framework for process management of research data; The ontology will allow re-users to understand the tasks and procedures performed with a common protocol. It may help to improve the data curation workflow when accepting research data from different fields.

## Ontology development

### Motivation

Our analysis of the actual data curation activities in multiple fields revealed that the data curation process has a structure consisting of Input-Output objects, hierarchical relationships among activities, and staffing. Since these relationships are complicated, it is not easy to represent the relationships in a simple tabular form; We need some method for representing the relationships adequately.

### Method

We have developed "Data curation process ontology" that collects and structures knowledge to represent the data curation activity's structure.

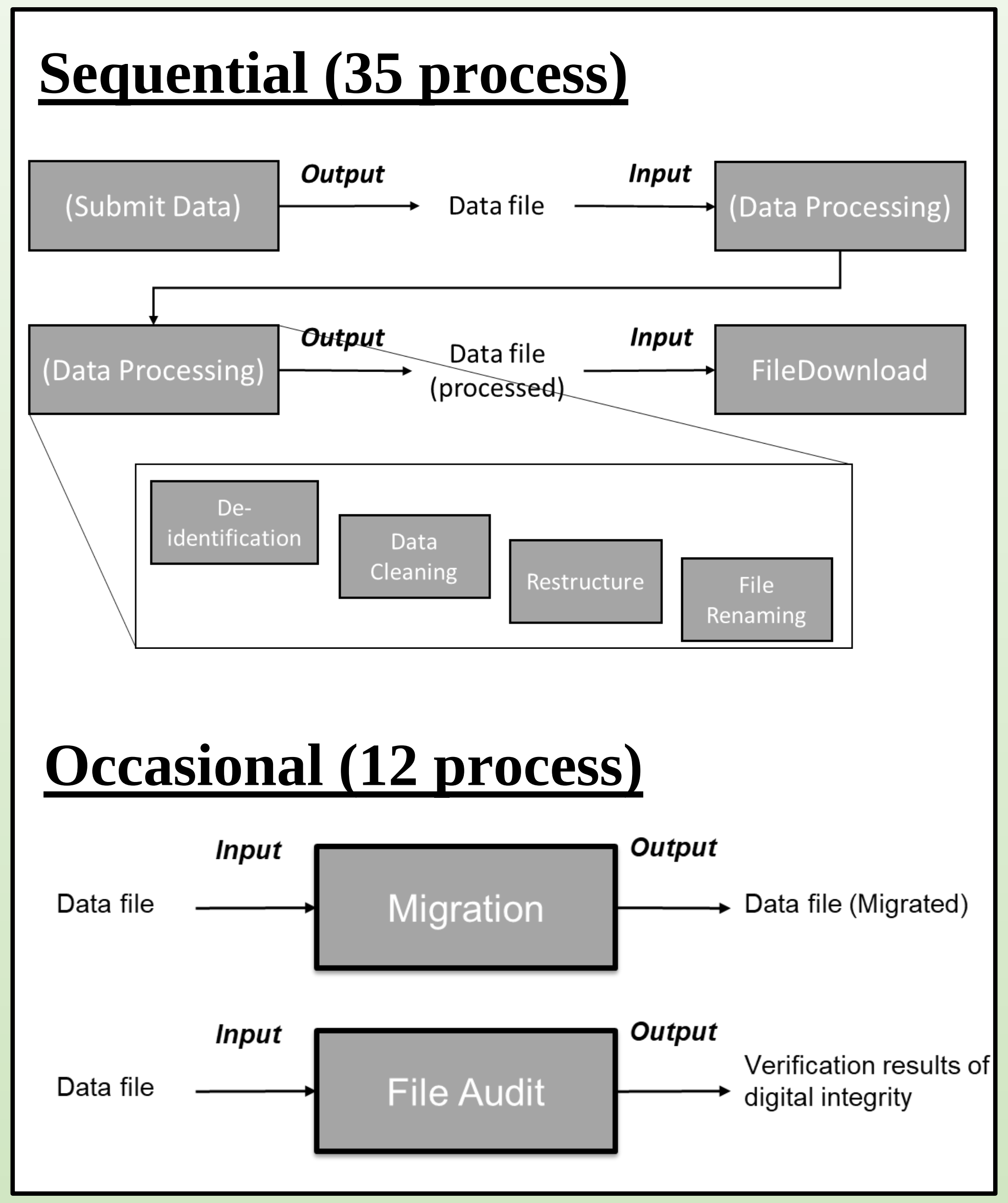
**Data Curation Activities**

Data curation activities can be arranged around five steps of data curation life-cycle: Ingest, Appraise (Accept), Curate, Access, and Preserve.

**Ingest**

- Authentication: The process of confirming the identity of a person, generally the depositor, who is contributing data to the data repository. (e.g., password authentication or authorization via digital signature). Used for tracking provenance of the data files.
- Chain of custody: Intentional recording of provenance metadata of the files (e.g., metadata about who created the file, when it was last edited, etc.) in order to preserve file authenticity when data are transferred to third-parties.
- Deposit agreement: The certification by the data author (or depositor) that the data conform to all policies and conditions (e.g., do not violate any legal restrictions placed on the data) and are fit for deposit into the repository. A deposit agreement may also include rights transfer to the repository for ongoing stewardship.
- Documentation: Information describing any necessary information to use and understand the data. Documentation may be structured (e.g., a code book) or unstructured (e.g., a plain text "Readme" file).
- File validation: A computational process to ensure that the intended data transfer to a repository was perfect and complete using means such as generating and validating file

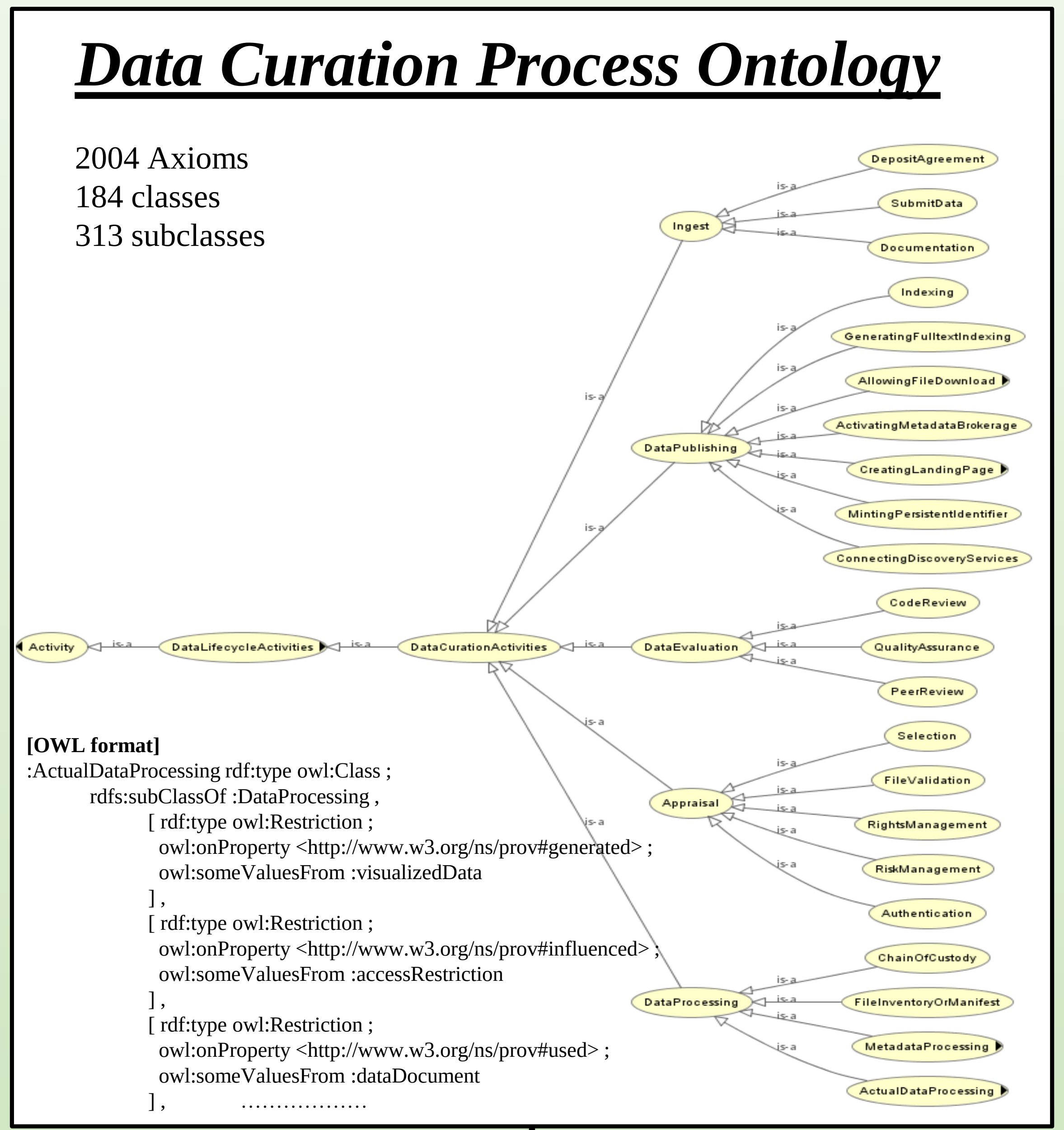
**Analyzed →**



**Sequential (35 process)**

**Occasional (12 process)**

**Formalized →**



***Data Curation Process Ontology***

2004 Axioms
184 classes
313 subclasses

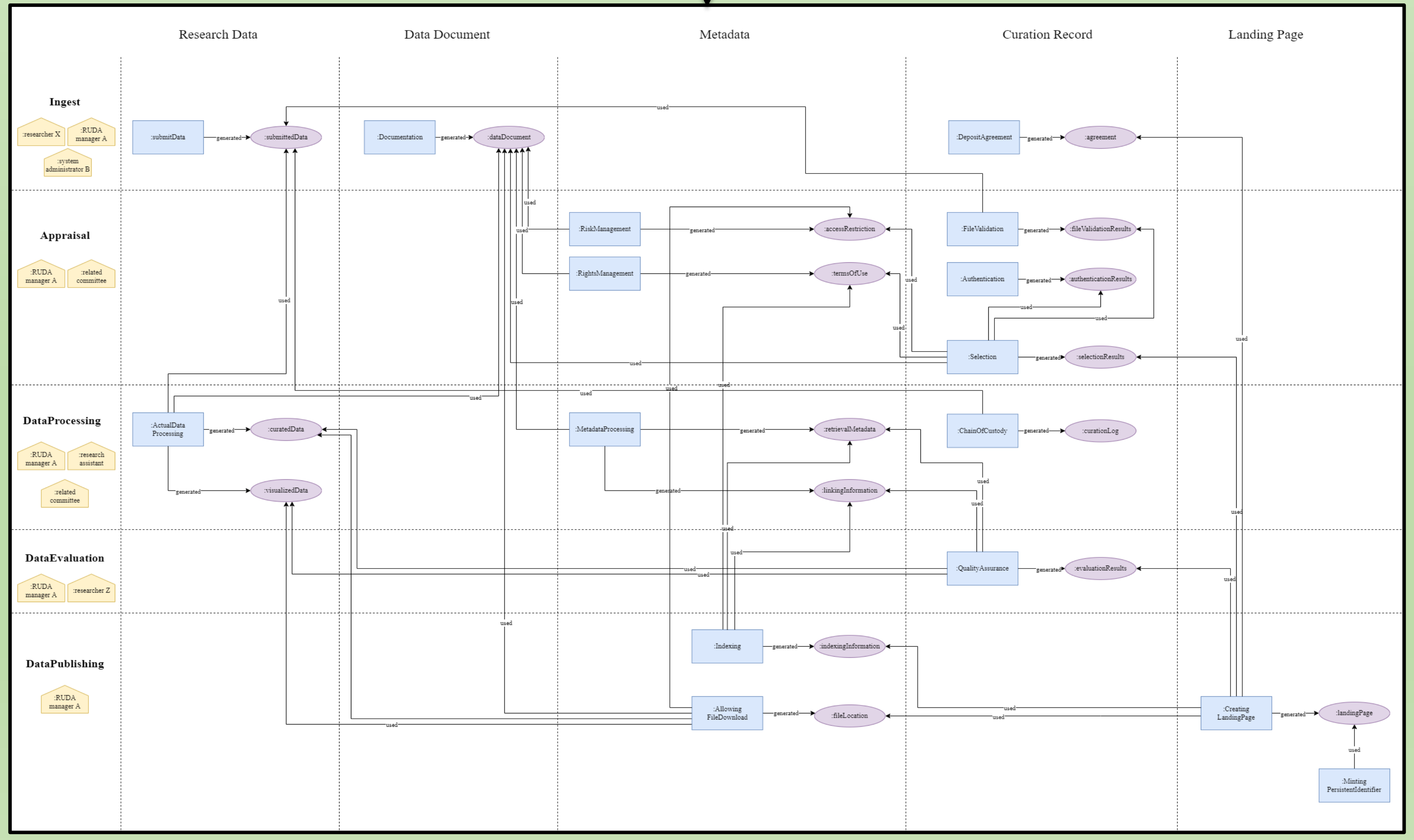
**[OWL format]**

```
:ActualDataProcessing rdf:type owl:Class ;
        rdfs:subClassOf :DataProcessing ,
                [ rdf:type owl:Restriction ;
                  owl:onProperty <http://www.w3.org/ns/prov#generated> ;
                  owl:someValuesFrom :visualizedData
                ] ,
                [ rdf:type owl:Restriction ;
                  owl:onProperty <http://www.w3.org/ns/prov#influenced> ;
                  owl:someValuesFrom :accessRestriction
                ] ,
                [ rdf:type owl:Restriction ;
                  owl:onProperty <http://www.w3.org/ns/prov#used> ;
                  owl:someValuesFrom :dataDocument
                ] ,          ………………
```

https://purl.archive.org/curation-ontology

## Ontology utilization

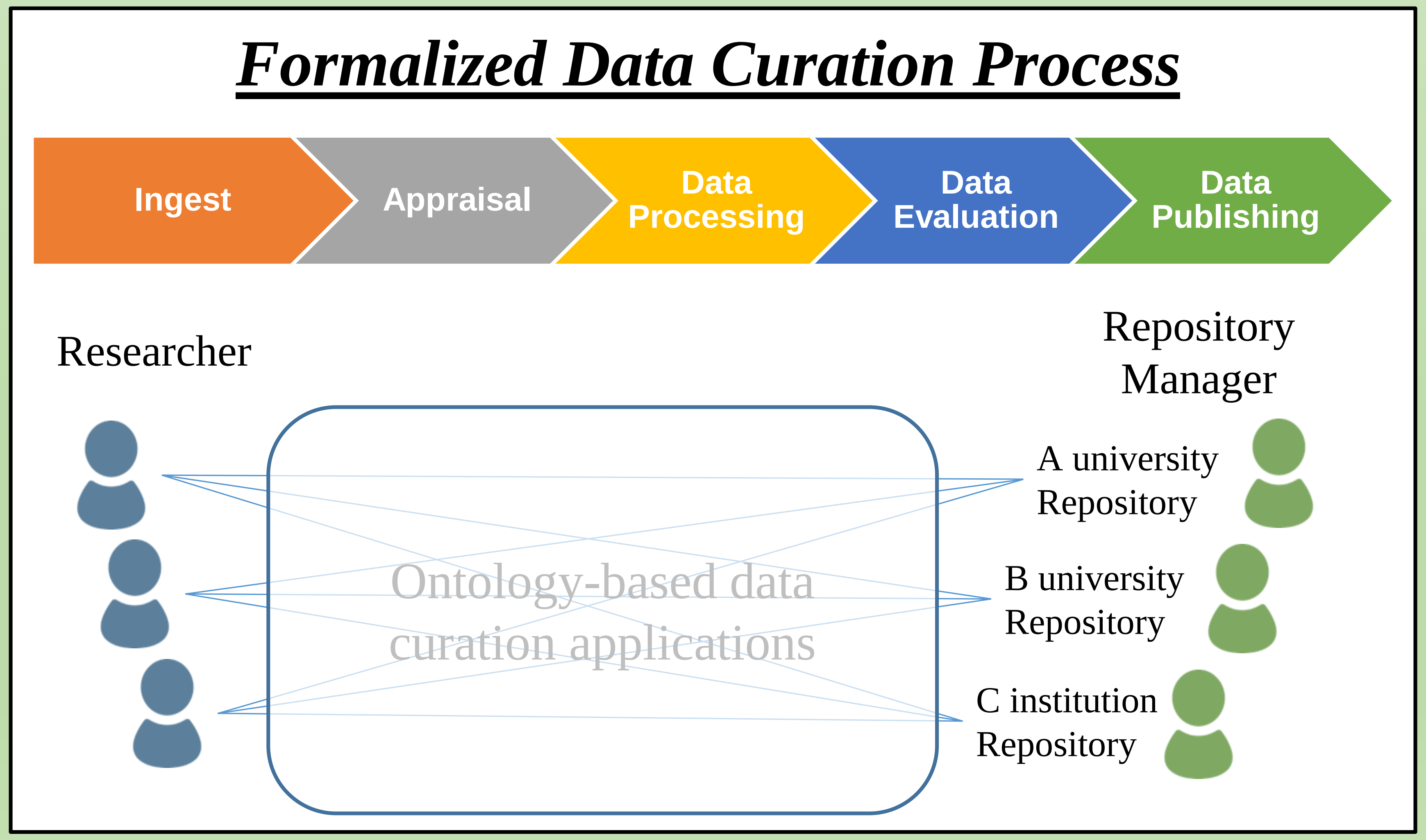
### Case study



By using this ontology, we represented the actual data curation processes in multiple fields as a diagram. The diagram is represented by a 5x5 matrix. The columns consist of five typical Entities and the rows consist of the categories adopted by the ontology.

**Visualized**

**Will be implemented →**



***Formalized Data Curation Process***

The ontology provides a well-formalized structure for the data curation activity and may function as a framework for process management of research data. Furthermore, combining different field-specific applications in each field will be possible.

### References

Johnston, L. R., Carlson, J., Hudson-Vitale, C., Imker, H., Kozlowski, W., Olendorf, R., & Stewart, C. (2016). Definitions of Data Curation Activities used by the Data Curation Network. University of Minnesota Digital Conservancy. https://hdl.handle.net/11299/188638

Lebo, T., Sahoo, S.S., McGuinness, D.L., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., & Zhao, J. (2013). PROV-O: The PROV Ontology. https://www.w3.org/TR/prov-o/

Minamiyama, Y. Data Curation Process Ontology [Data set] https://purl.archive.org/curation-ontology (accessed 2022-05-18)

Yasuyuki Minamiyama, Hideaki Takeda, Masaharu Hayashi, Makoto Asaoka, and Kazutsuna Yamaji
National Institute of Informatics, Japan. Contact: minamiyama@nii.ac.jp