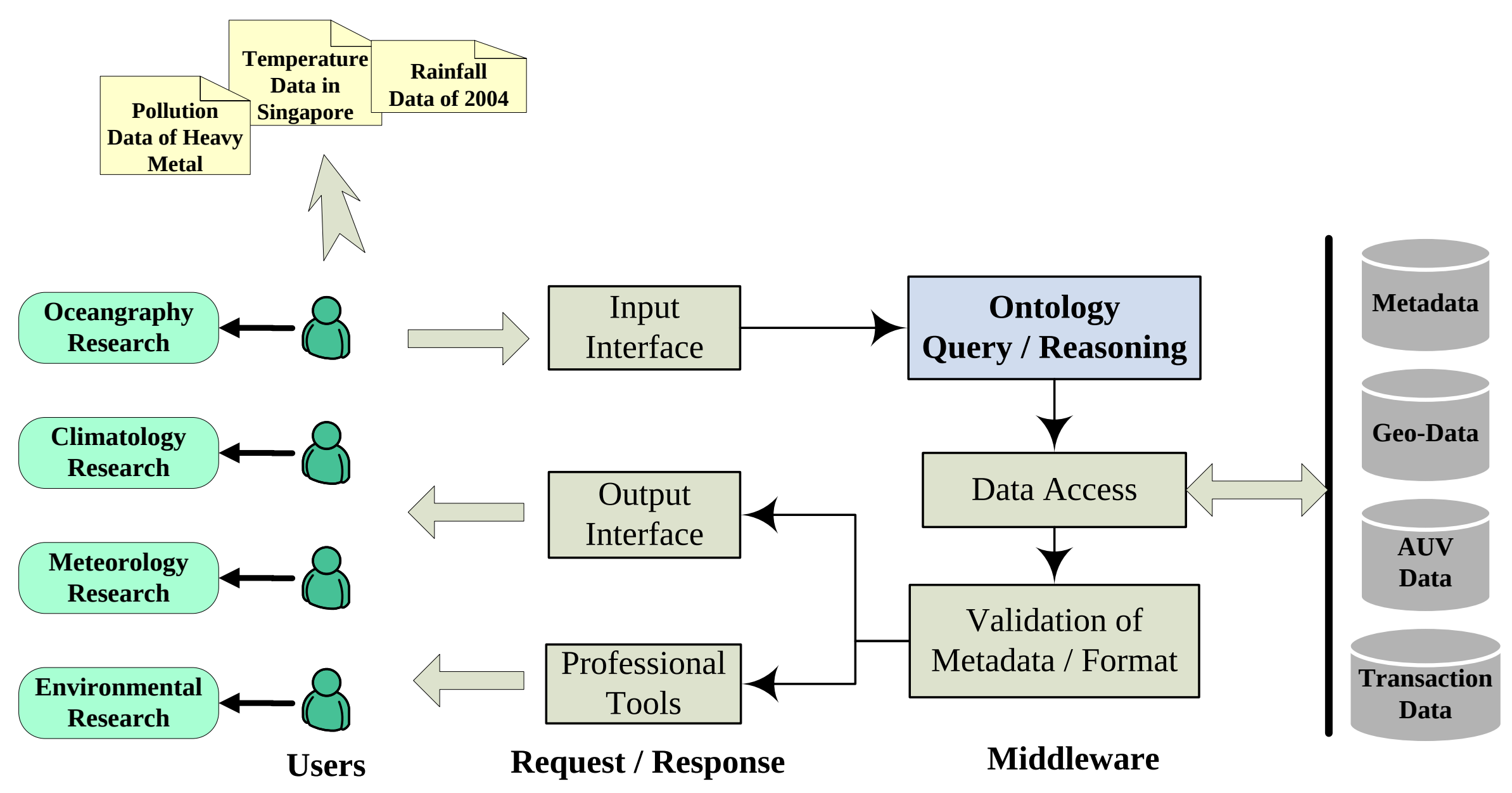
Ontology-based Environmental Data Exchange and Integration

Liang Yu (Post-Doc, NUS), **Byounghyun Yoo** (Research Scientist, SMART/MIT), **Chen-Chieh Feng** (PI, NUS), **V. Judson Harward** (PI, MIT)

geoly@nus.edu.sg, byoo@mit.edu, geofcc@nus.edu.sg, jud@mit.edu

1. Motivation and Goal

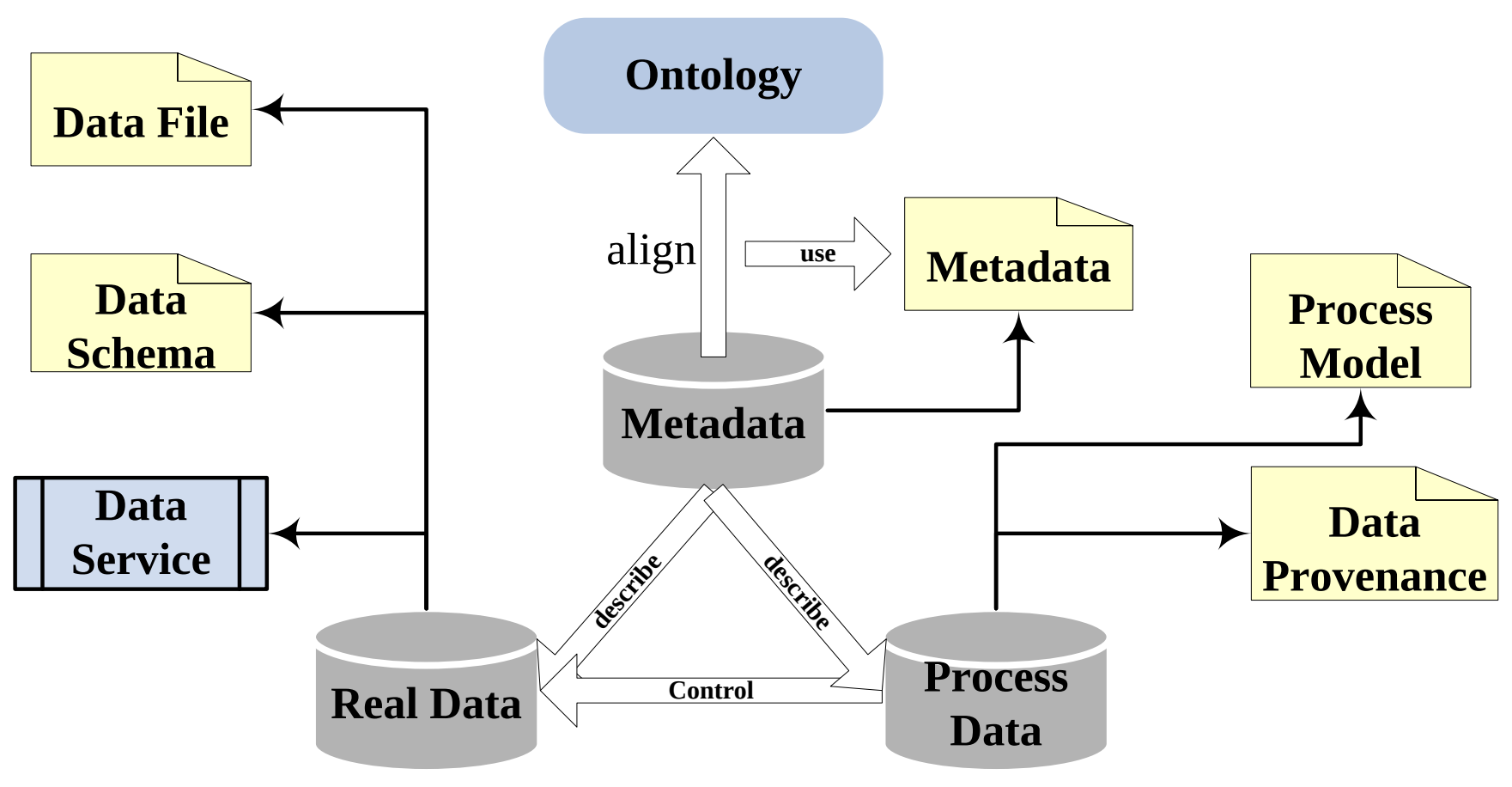


Research groups in CENSAM come from different backgrounds, use different tools, and generate huge amount of data that differ from formats, standard vocabularies, units, reference systems, and so on.

It is a challenging task for different research groups to identify the data that are not produced in-house, let alone integrating the data. This work aims to design a semantic-based data identification and integration system and to develop a tool for easy access of data from different sources. The figure shows the data access process from a scientific researcher’s point of view. The user comes up with some concepts and their restrictions, which are translated to a specific query request by the ontology reasoning component, and then the matched data were returned coupled with metadata.

2. Tasks

● Data production and storage. It includes raw data from sensors and processed data from simulation, assimilation, and data processing. In order to provide mechanisms for data discovery by researchers as well as data access for authorized users, the data is managed with a database management system.

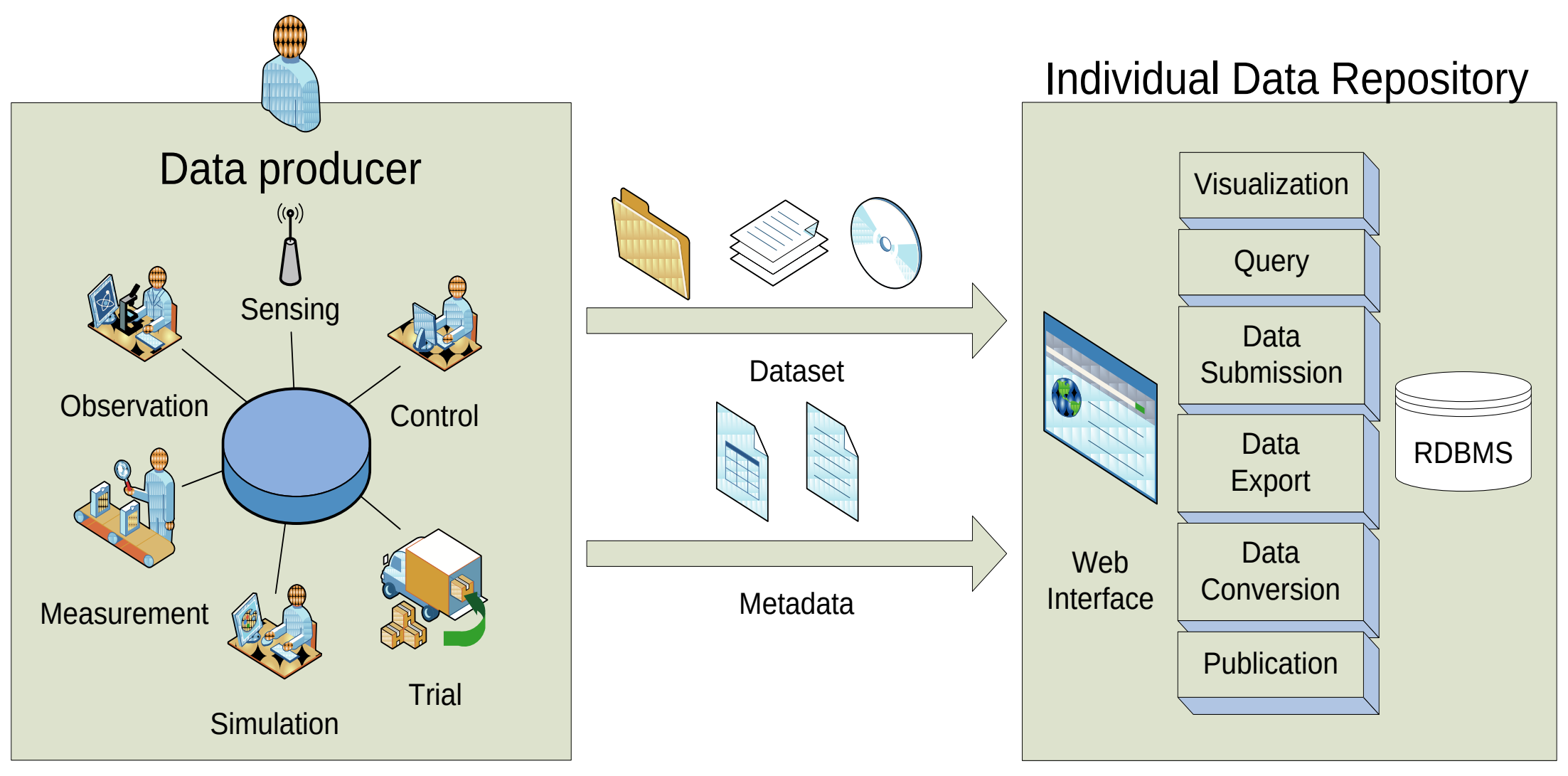


● Metadata and process model development. Metadata describes the information for every single dataset. Process model describes a system how to take in input data and generate output data. The physical process model is typically a sensor or a sensor group, while non-physical process model is typically a work flow of multiple data processing chains.

● Ontology design and alignment. Develop an ontology containing the concepts across our research domains and specify the relations between each pair of concepts necessary for data integration. These concepts are used to annotate different parts of the data, e.g., the “Keywords” and “Entity and Attributes” from metadata and the “Input” and “Output” from a SensorML-based document. The alignments facilitate the search and the conversion process with the help of reasoning.

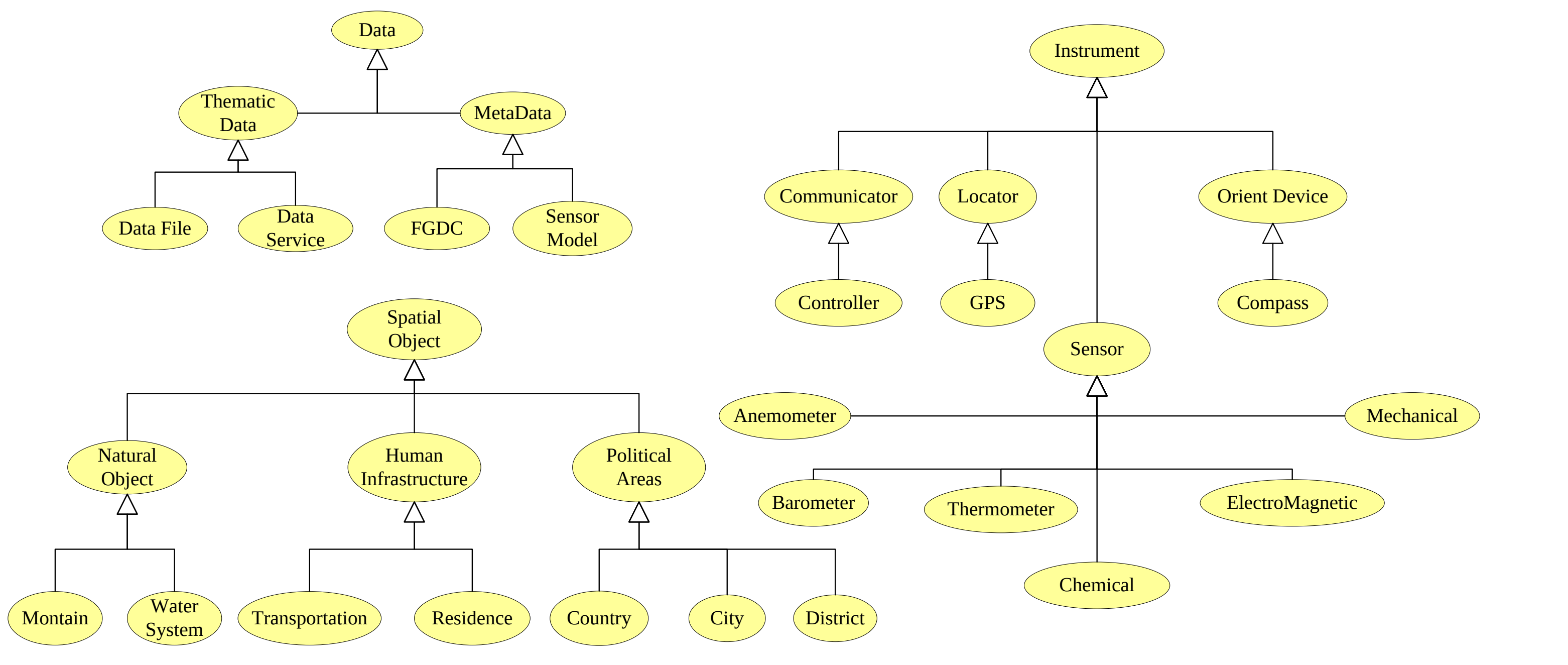
3. Data Storage and Publication

Domain knowledge and metadata can be captured from data producers and domain experts. This information is used for designing a database system for individual projects. A web-based user interface of data repository is designed to provide these capabilities:



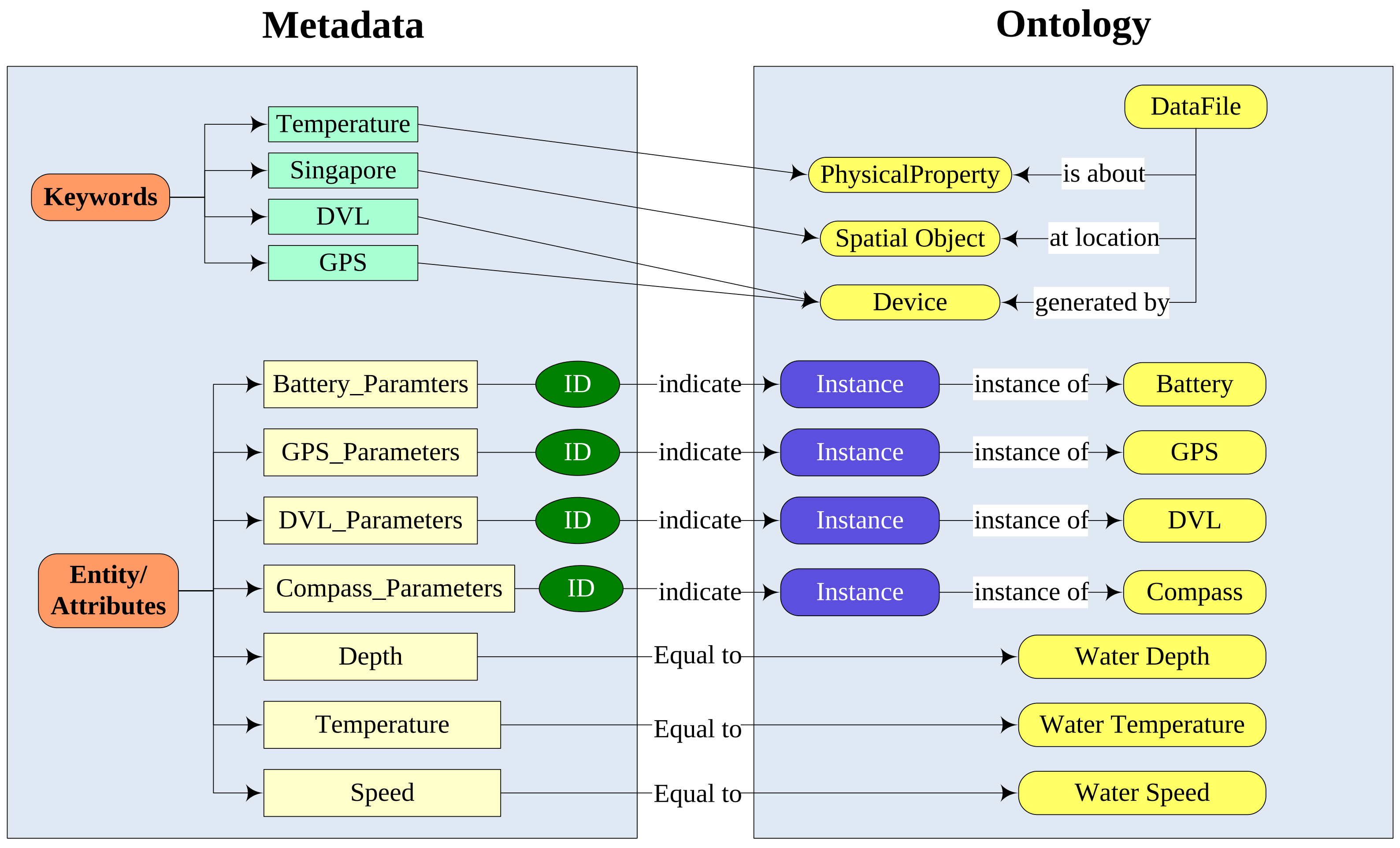
- Upload dataset into the backend database
- Mechanism to parse and store the dataset
- Data query interface
- Data export interface that support popular file formats such as Excel, Word, XML, CSV, and PDF
- Web-based visualization of dataset

4. Ontology Design



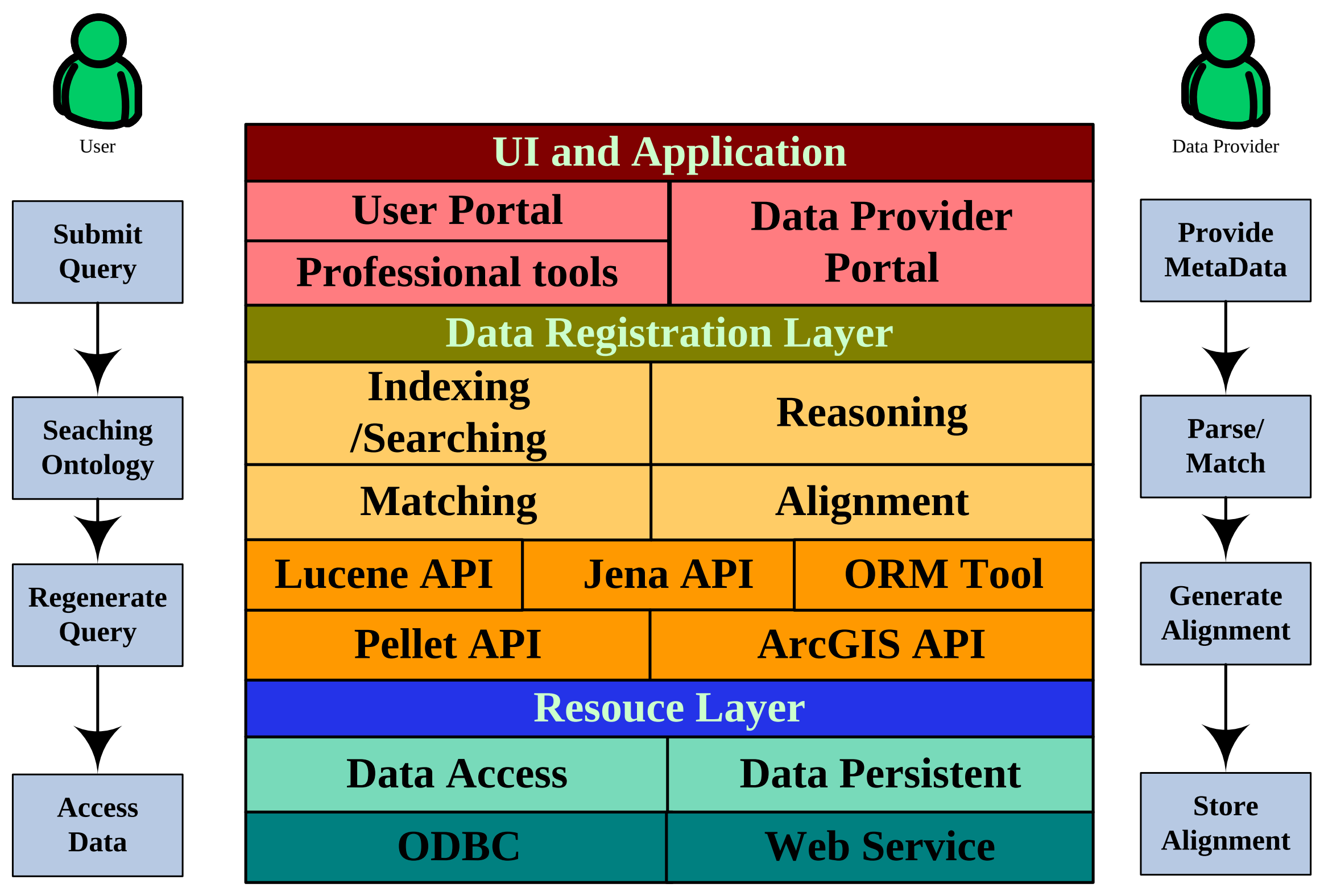
We build our ontology on the basis of formal ontologies and standards, which include SWEET, CSDGM, and SensorML. To improve the usability of SWEET we imported a portion of the concepts and at the same time extended it by adding new concepts to ontology, which might exist in vocabularies from other CSDGM or SensorML and specifying the relations between concepts.

5. Ontology Alignment



The information in the metadata (e.g. keywords, entities and attributes, and spatial domains) were extracted and aligned to ontology either manually or automatically, depending on the vocabularies it used. The information can be aligned to a concept, an instance of a concept, or a property of concept in the ontology. These alignments are recorded in the system and can be used in further searching and integration.

6. System Architecture



There are basically three layers in the system. The middle layer use the ontology and other APIs to perform the original information processing, which contains two paths:

● For a data user, it is much easier and straightforward to search by concepts rather than by looking into the database. He/she uses the ontology to create a request, which will be processed by the reasoning engine and attached with more semantic meanings (e.g. more concepts and restrictions). Using the alignment information, the original query will then be translated to different forms suitable for querying different data sources.

● For a data provider, he/she will provide sufficient metadata according to uniform standards. Each of the metadata elements can be directly mapped to ontology concepts or some other standard vocabularies like those in GCMD to be recognized and aligned automatically. Users can also add more alignments or alter existing ones manually. Both of the original metadata and generated alignment are stored in the database for future applications.

7. Future Work

1. Evolve the ontology to satisfy the needs of different groups and investigate the way to evolve existing alignment information meanwhile.

2. Investigate data format conversion and the application integration to support both data and service sharing, e.g., users can click a quick link to view a searched spatial dataset via an online visualization service. In this process, data should be organized and converted to a format suitable to be the input of the service. User can also choose to download them in specific format and use local tools to handle them.

3. Develop tools to help data providers automate the data publication and decrease their data maintenance load, such as the load to edit and validate metadata.

4. Extend the reasoning capability of the system by introducing spatial concepts such as *near*, *far*, and *neighbor* and perform ontology-based spatial computation. Spatial computation is supported by the functions in ArcGIS API. Integrating spatial computation with ontology components especially the reasoning functions would improve significantly the data search capability of the system.

Acknowledgements

The research described in this project was funded in whole or in part by the Singapore National Research Foundation (NRF) through the Singapore-MIT Alliance for Research and Technology (SMART) Center for Environmental Sensing and Modeling (CENSAM).