Fostering Resource Integration: EarthCube Resource Registry

Rebecca Koskela[1], Stephen Richard[2], Ilya Zaslavsky[3], Anna Kelbert[4], and Ruth Duerr[5]

[1]University of New Mexico
[2]U. S. Geoscience Information Network (USGIN)
[3]San Diego Super Computer Center
[4]USGS Geological Hazards Science Center
[5]Ronin Institute for Independent Scholarship

November 21, 2022

Abstract

The EarthCube Technology & Architecture Committee formed a Resource Registry Working Group (WG) to develop a framework for a registry of EarthCube (EC) resources, enabling users to discover scientific and technical resources (software, tools, vocabularies, etc.) that are relevant to their research. The registry will promote EC investments, reduce time to science, help enable interdisciplinary research, more clearly define what is EC, and provide a vehicle for tool and software producers to notify the community about new products, increase visibility, and gain recognition. A primary requirement is to enable systematic description of EarthCube computational resources in terms of their functionality and interfaces for utilization, to enable users to identify components that can work together in integrated workflows. This requires understanding the specifics of how a software component communicates—both the messaging protocol, and the syntax and semantics of information formats getting data into and out of a component. This registry would work in conjunction with schema.org dataset descriptions being developed by the community to streamline linkage of data and software components for research workflows. The WG created definitions for a set of resources to include in a first iteration of the registry, and a set of properties that should be specified for all resources, as well as properties specific to particular resource types. The suggested resource types are: Software, Interface/API, Interchange format, Dataset, Repository, Service, Platform, Vocabulary/ontology/Information model, Specification, Catalog/registry, and Use Case. Dataset and Use Case resources registration is out of scope for the WG project, to be handled separately. Elaboration of this registry is in the workplan for EarthCube, with the goal maximum reuse of existing vocabularies and technology and compatibility with related registry activities.

Posted on 21 Nov 2022 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.1002/essoar.10500455.1 — This a preprint and has not been peer reviewed. Data may be preliminary.

IN23B-0776

Fostering resource integration: EarthCube Resource Registry

Rebecca Koskela, University of New Mexico; **Stephen M. Richard**, LDEO, Columbia University and US Geoscience Information Network (USGIN); **Ilya Zaslavsky**, San Diego Supercomputer Center; **Anna Kelbert**, USGS Geological Hazards Science Center; and **Ruth Duerr**, Ronin Institute for Independent Scholarship

What is the EarthCube Resource Registry?

A database of cyberinfrastructure resources with documentation focused on enabling EarthCube users and developers find, understand, get, and use those resources to increase research productivity

Resource focus: applications, reusable code components, ontologies, vocabularies, specifications (extend p418 scope)

Why?

- Improve discovery of usable resources
- Enable seamless connection from discovery environment to working with the data or software.

Some usage scenarios:

- Identification of resources
- Systematic documentation of resource characteristics
- Identify gaps or duplicate resources
- Find tools, APIs, or data that can work with a given resource
- Support maturity assessment

The Registry should:

- Help researchers connect multiple data types and resources to address a specific research problem;
- Enable developers to learn about components they can reuse to increase development efficiency
- Enable discovery of components with a particular functionality, that can be used in an existing research workflow
- Provide a platform for resource producers to inform the community about products.
- Answer the question: “What has EC produced that is of use to my science?”

Acknowledgements: Work to refine the registry concept and draft model was done at a workgroup meeting held in March, 2018, funded by the NSF EarthCube ESSO office; participants were Rebecca Koskela, Anna Kelbert, Simon Goring, Ruth Duerr, Stephen Richard, and Ilya Zaslavsky

Resource Types

Resource types and documentation are focused on providing information necessary to understand how resources can be used together

Software	**Service**
Interface/API	**Vocabulary/ontology/Information model**
Interchange format	**Specification**
Dataset	**Catalog/registry**
Repository	**Platform**
	Use Case

Resource type properties

- Specify the characteristics of the resource useful for finding and determining how to use
- Resource types are different if they have different properties

Property examples

Properties that apply to any resource

- name
- system/project
- description
- identifier
- URL to user-readable page
- primary publication

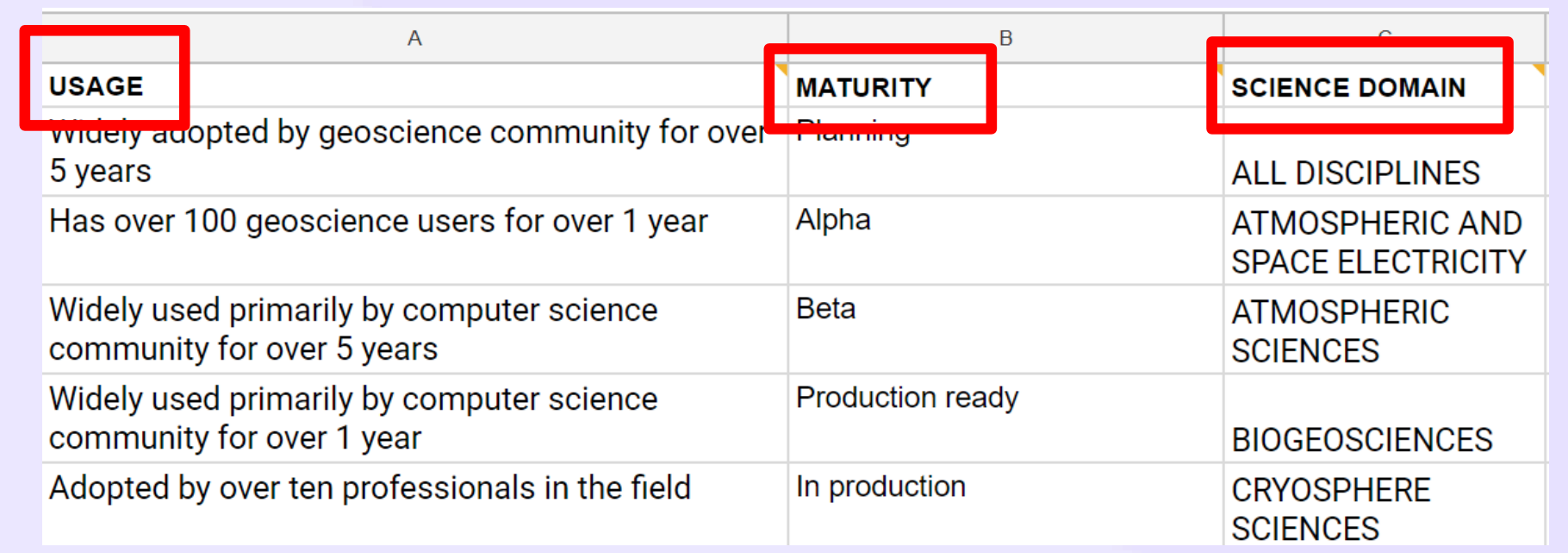
….

Properties that apply to specific resource types

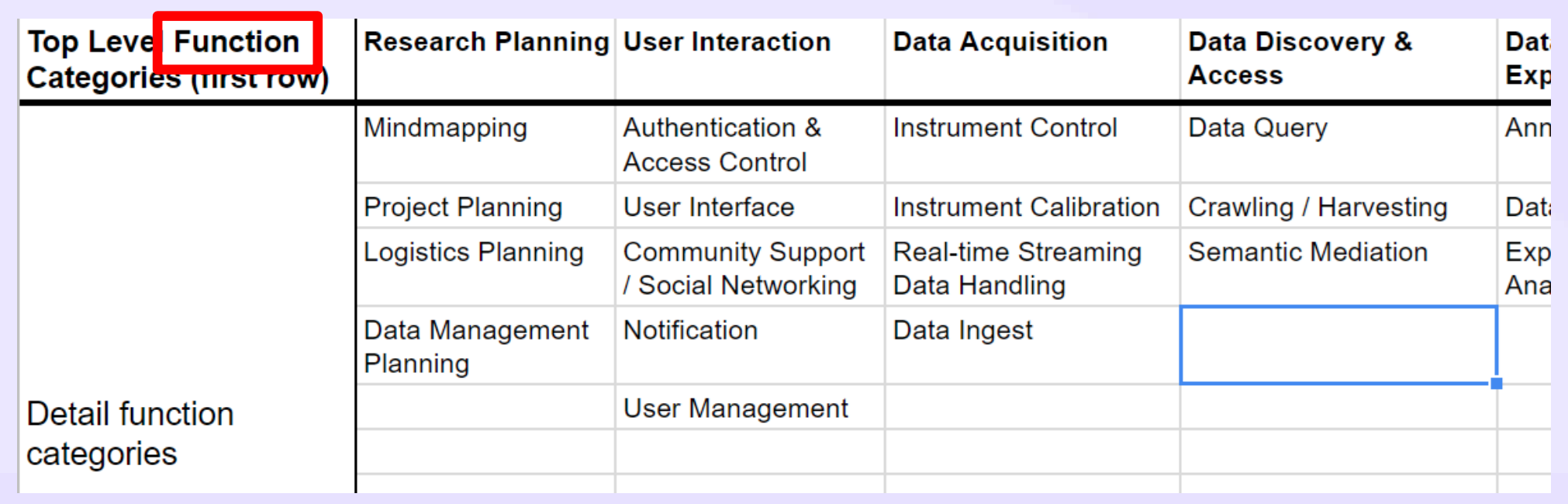
- Function
- Model/Algorithm Implemented
- Interface
- Input file format
- Output file format

…..

Some properties need controlled vocabularies



USAGE	MATURITY	SCIENCE DOMAIN
Widely adopted by geoscience community for over 5 years	Planning	ALL DISCIPLINES
Has over 100 geoscience users for over 1 year	Alpha	ATMOSPHERIC AND SPACE ELECTRICITY
Widely used primarily by computer science community for over 5 years	Beta	ATMOSPHERIC SCIENCES
Widely used primarily by computer science community for over 1 year	Production ready	BIOGEOSCIENCES
Adopted by over ten professionals in the field	In production	CRYOSPHERE SCIENCES



Top Level Function Categories (first row)	Research Planning	User Interaction	Data Acquisition	Data Discovery & Access	Dat Exp
	Mindmapping	Authentication & Access Control	Instrument Control	Data Query	Ann
	Project Planning	User Interface	Instrument Calibration	Crawling / Harvesting	Dat
	Logistics Planning	Community Support / Social Networking	Real-time Streaming Data Handling	Semantic Mediation	Exp Ana
	Data Management Planning	Notification	Data Ingest		
Detail function categories		User Management			

draft examples from the workgroup

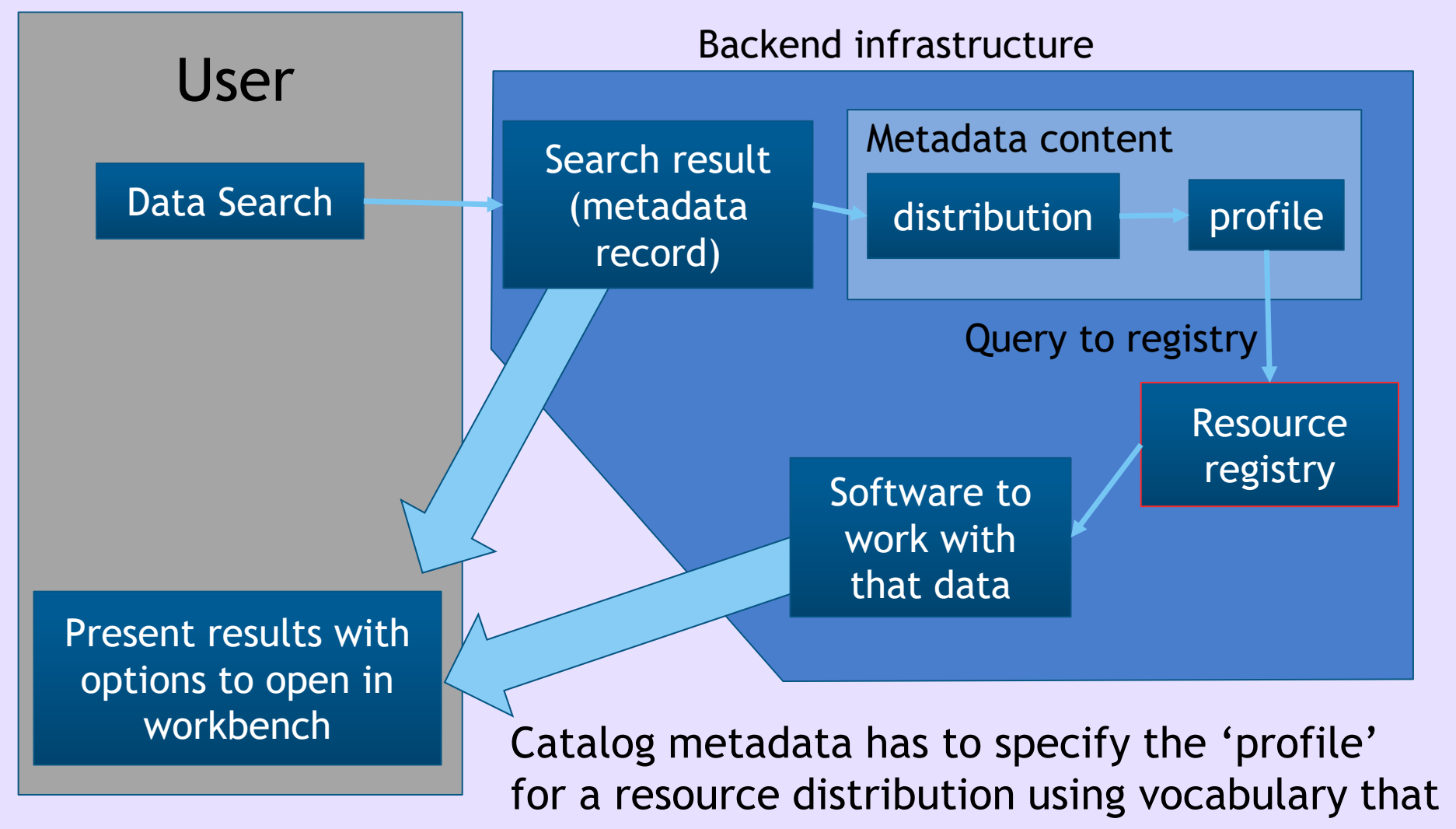
Not starting from scratch!

- Earlier EC resource inventories
  - High-level geoscience infrastructure resources from EC Roadmaps, workshops, earlier projects
  - Domain resource catalogs compiled by RCNs (C4P, SEN, ECOGEO, CRESCYNT)
  - EC tools inventory
- Vocabularies and schemas from EC projects
  - CSDMS Names
  - Anna Kelbert’s categorization
  - Recommendations for resource description from ESSO P418
  - Categorizations used in earlier inventories and the GEAR model for EC inventory items
- Standards and best practices from other registries (e.g. NITRC, DiRT, GEOSS, INSPIRE, OntoSoft)

Resource Types Detail

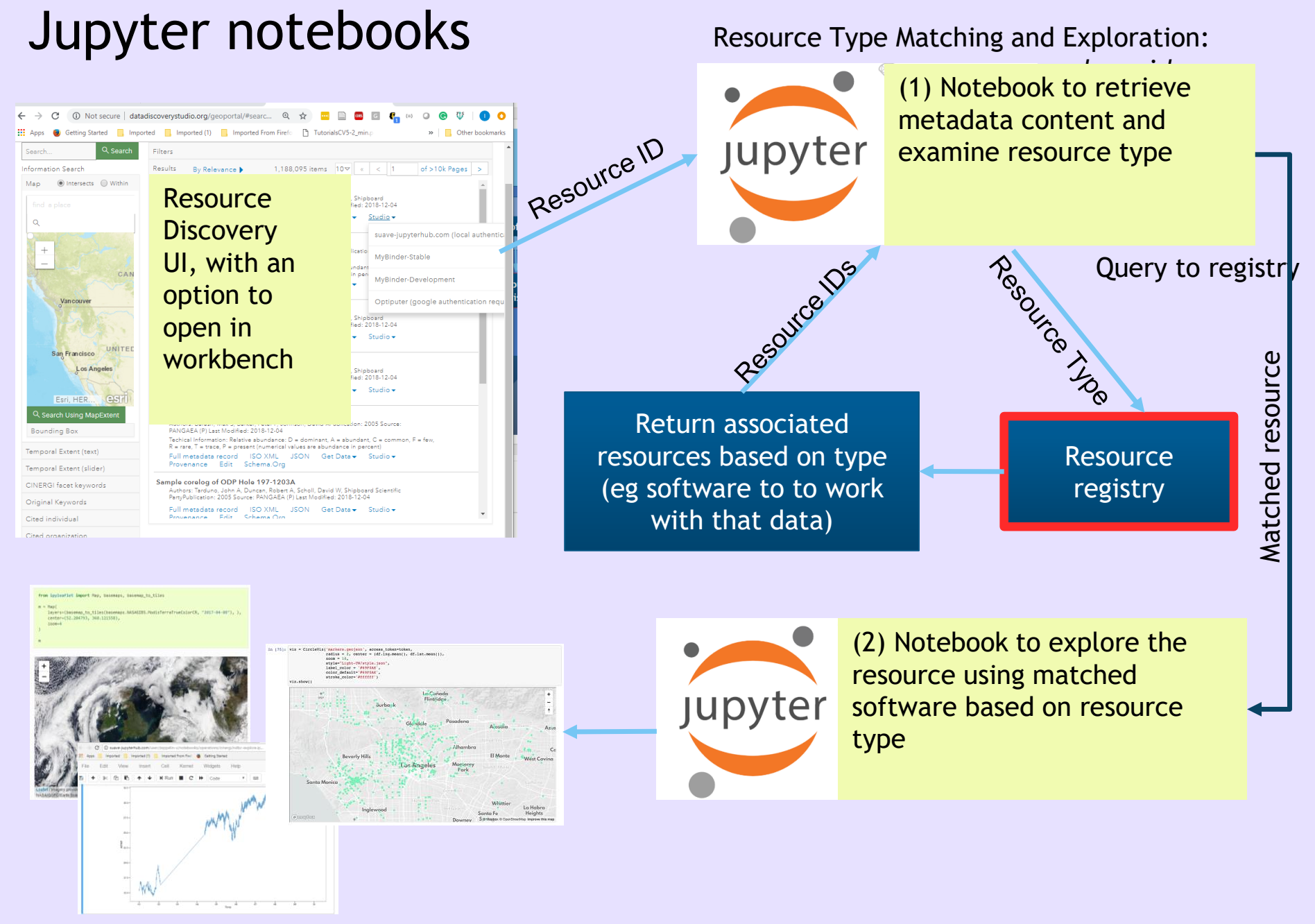
Resource type	Scope Note	Properties
Software	a packaged set of instructions that can be executed by a machine to perform one or more functions	Function, Model/Algorithm implemented, Interface, input file format, output file format, programming language, execution environment
Interface/API	Specification of a set of operations, messages to invoke the operations, inputs necessary to execute an expected output content and format	Communication protocol (http...), operations, message(content), schema (interchange format)
Interchange format	Specification of a serialization scheme (file format) that implements some information model (schema and vocabulary) to communicate information between agents	Information model, serialization format, vocabularies
Dataset	a collection of data items unified by some criteria (authorship, subject, scope, extent...). A kind of Collection that contains data items (See Yamz http://www.yamz.net/term=h1043)	Information model, vocabularies, representation formats [note-- we don’t intend to populate all datasets in this registry, but we we want to specify the properties that need to be included]
Repository	a storage system in which objects may be stored for subsequent access or retrieval (generalize from Kahn and Wilensky, 1995, http://www.cnri.reston.va.us/k-w.html)	Scope (content types), interfaces
Service	A computation performed by a software entity on one side of an interface in response to a request made by an agent on the other side of the interface. A collection of operations, accessible through an interface, that allows an agent to evoke a behavior of value to the user. Source: ISO 19119	Dataset(content) offered, functionality offered, interfaces
Vocabulary/ontology /Information model	A specification of concepts, and optionally relationships representing a conceptualization of some domain of discourse	Scope, interfaces, representation format
Specification	A document that describes the technical characteristics of an artifact or practice, possibly including a description of what it should do, or an explicit set of requirements that it must satisfy. http://en.wikipedia.org/wiki/Specification. e.g. interoperability agreement, identifier scheme	Scope, File Format
Catalog/registry	A curated collection of descriptions of resources, accessible through one or more interfaces	Scope, interfaces
Platform	A composite software entity that enables execution of a variety of tools e.g. MatLab, ArcGIS	Extension programming language; scope, interface, function
Use Case	A specification of a scenario for a work item with some specific context and goal.	Interdisciplinarity level {low, medium, high}; Science Theme;

Role in EC architecture: Data Discovery

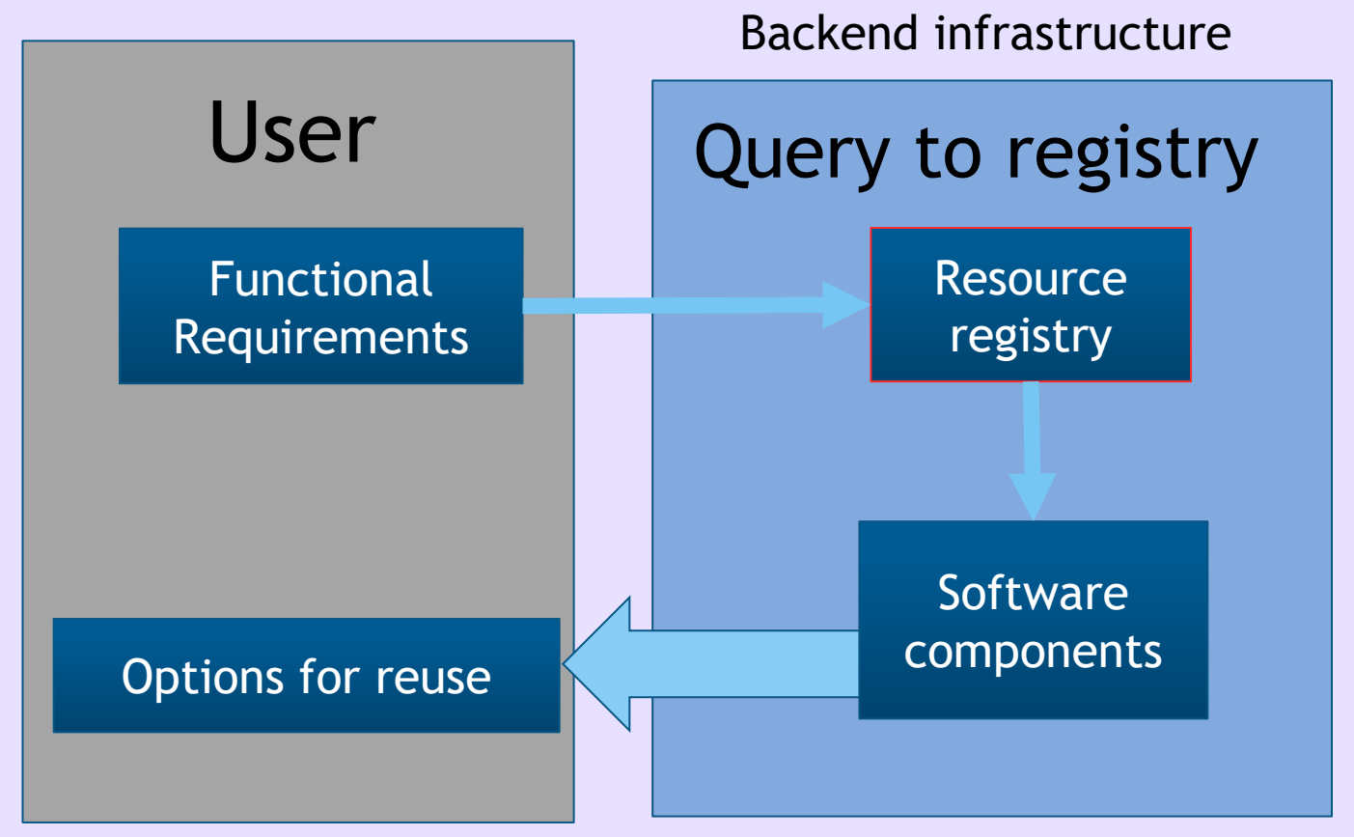

User
Data Search
Present results with options to open in workbench
Backend infrastructure
Search result (metadata record)
Metadata content
distribution
profile
Query to registry
Resource registry
Software to work with that data

Catalog metadata has to specify the ‘profile’ for a resource distribution using vocabulary that enables query to registry to find compatible resources

Example scenario, DataDiscoveryStudio and Jupyter notebooks


Resource Type Matching and Exploration:
Resource Discovery UI, with an option to open in workbench
Resource ID
jupyter
(1) Notebook to retrieve metadata content and examine resource type
Query to registry
Resource Type
Resource IDs
Return associated resources based on type (eg software to to work with that data)
Resource registry
Matched resource
jupyter
(2) Notebook to explore the resource using matched software based on resource type

Role in EC architecture: software development



Have to match function, as well as determine if input and output interfaces are compatible with use scenario

Find reusable component that implements a particular function

Work plan

- Select a representative set of EarthCube resources to document, focusing on resources listed in the EarthCube tool inventory web page.
- Assess requirements and refine information model and vocabularies proposed by the March, 2018 Registry Working group workshop
- Compile descriptions in spreadsheets (e.g. http://bit.ly/2M2shmF)
- Develop an RDF document format to link the acquisition interface with the back-end database (JSON-LD, base on patterns from p418 and other existing vocabularies)
- Select and deploy database
- Transform tabular compilation to RDF, and load in database
- Develop demonstration queries and documentation