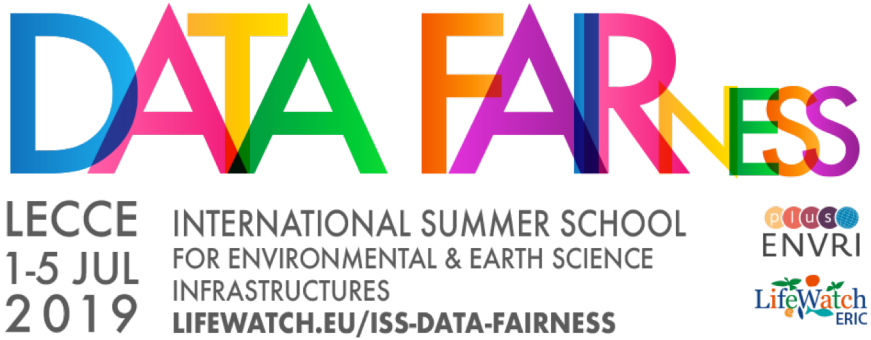


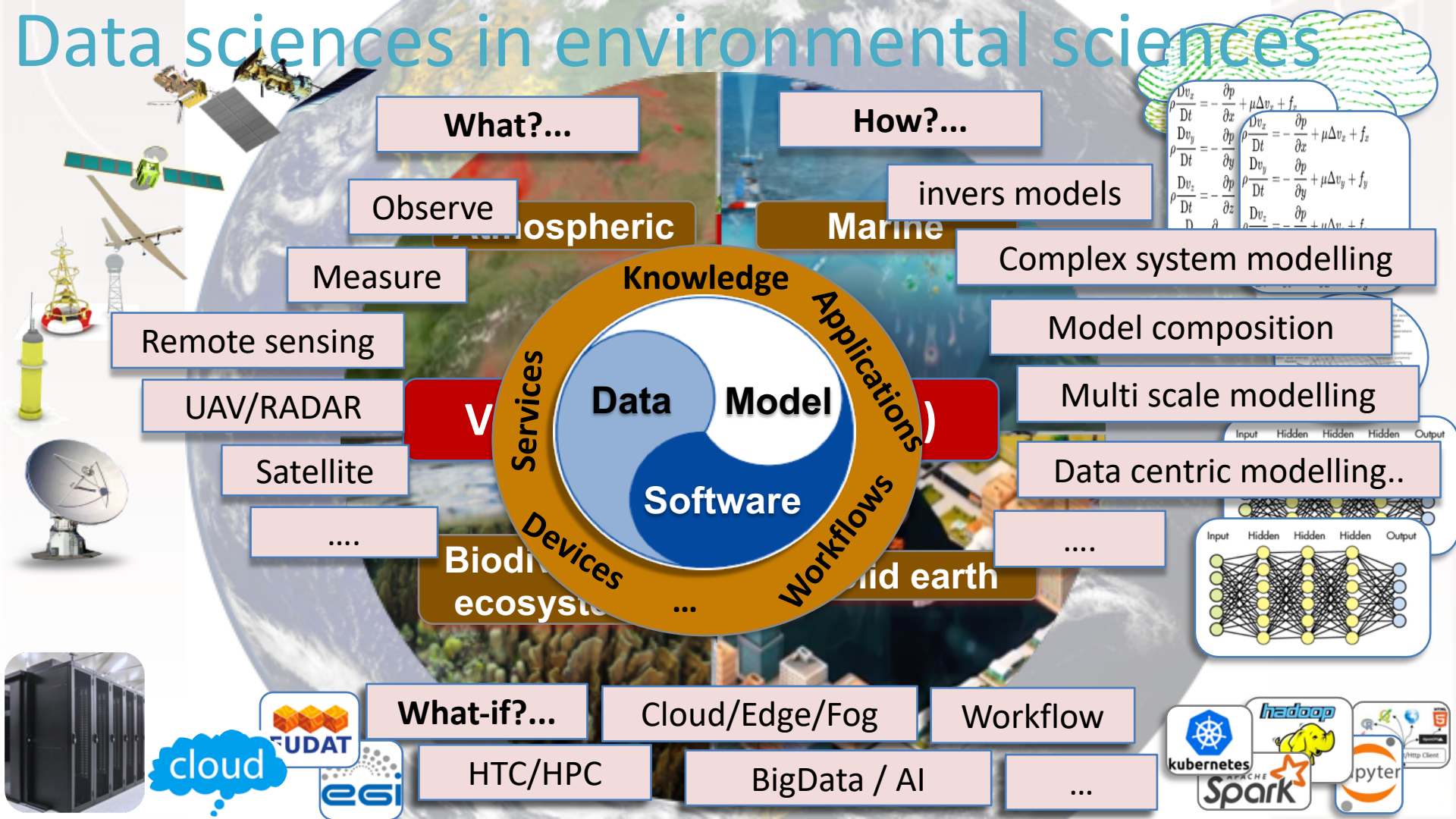
Catalogue and cataloguing

Zhiming Zhao

z.zhao@uva.nl

UNIVERSITY OF AMSTERDAM

Data sciences in environmental sciences




SIOS
EISCAT
ICOS
ESONET
Euro Argo
Cannot find
ACTRIS
IAGOS
EUFAR
EuroGOOS
European Global Ocean Observing System
Eurofleets
LTER
FixO3
AnaEE
JERICO
LifeWatch
is-enes
SeaDataNet
elixir
EMBRC
EUROPEAN MARINE BIOLOGICAL RESOURCE CENTRE
EPOS
emso
....
An effective catalogue is needed!

Outline

- Catalogue: from different points
- How to catalogue?
- Examples
- Discussion

What is a catalogue?
- in a mall

APPAREL - ATHLETIC WEAR
B7 Foot Locker 905.578.5063
A11 Lids 905.561.1111
F1 SportChek 905.560.9371

APPAREL - CHILDREN'S
C8 The Children's Place 905.573.3808

APPAREL - LADIES
E5 Alia n TanJay 905.561.1022
C6 Anna Bella 905.560.1288
B1 Ardene 905.963.0309 x182
G3 Chelsea 905.561.4609
E6 Eclipse 905.578.1470
E4 Garage 905.560.4894
E2 La Senza 905.560.0190
E1 Le Chateau 905.573.8890
E8 Northern Reflections 905.573.2155
C7 Reitmans 905.560.4810
G7 Suzy Shier 905.561.4203
A15 Zaysha 905.578.4208

APPAREL - MEN'S
E10 International Clothiers . . . 905.573.7480
C16 Moores Clothing for Men . . 905.561.4265

APPAREL - UNISEX
C9 Bluenotes 905.560.4408
C6 Fotnwether/Stockhomme . . 289.965.8374
A9 Rock Zone 905.297.7056
F5 Urban Planet 905.578.2677

BOOKS & NEWS
E3 Coles 905.560.3784
D10 International News 905.561.1847
K17 Smokers Den 905.662.0928

CARDS, STATIONERY & GIFTS
A7 Carlton Cards 905.662.0217
B2 Spencer Gifts 905.560.4853

DEPARTMENT STORES
Sears 905.545.4741
F2 Winners 905.560.7366

ELECTRONICS, PHOTOGRAPHY & PHONES
C10 Bell 905.578.2140
D11 EB Games 905.662.8410
K12 Fido 905.560.8150
K2 In Touch 905.297.8452
A12 Koodo 905.560.6720
E7 Rogers 905.560.7027
D17 Tbooth 905.594.1206
D9 Telus 905.560.1739
A6 The Source 905.560.0269
K14 Virgin Mobile 905.573.7039
C8 Wind Mobile 289.448.5301
K6 Wirelesswave 905.561.0242
C13 Wireless Valley 905.578.0800
K15 WOW! mobile boutique . . 905.560.5758

FASHION ACCESSORIES
B9 Claire's 905.662.2592

FAST FOOD
A3 A&W 905.930.7999
D4 Jimmy the Greek 905.661.1000
D2 Kung Pao Wok 905.578.8882
D6 New York Fries 905.578.5020
D14 Orange Julius/DQ 905.560.7421
A2 Pizza Pizza 905.560.9104
D1 Subway 905.573.8836
D5 Thai Express 905.560.8788
D3 Tim Hortons 905.561.8534
K10 Tim Hortons Kiosk 905.560.8444

FINANCIAL SERVICES
B12 C.I.B.C Banking Centre . . 905.561.3600
D16 Continental Currency Exchange . 905.561.6100
B5 Liberty Tax 905.560.6890

FOOTWEAR
B7 Foot Locker 905.578.5063
B4 Payless Shoe Source 905.662.0832
C14 Quick Heel 905.560.7300
C4 SoftMoc 905.560.7483
G4 Stepps 905.930.8978

GENERAL MERCHANDISE & VARIETY
Dollarama 905.578.1740
A4 Showoose 905.578.5222

GROCERY STORES
Fortinos 905.561.5291

HEALTH/BEAUTY & PERSONAL CARE SERVICES
K9 Beauty 21 Brow Studio . . . . . N/A
C11 Dr. Jang (OPTOMETRIST) 905.561.1110
C1 Eastgate Dental Centre . . 905.560.2714
A10 CNC 905.561.5224
D13 Great Expectations 905.573.0770
B11 HairCrafters 905.573.9652
C13 Hearing Solutions 905.560.8585
C2 L'Amour Nails 905.578.6245
B3 Le Salon & Spa 905.560.2600
C10 LensCrafters 905.578.3292
C15 MasterCuts 905.662.7222
D19 Nutrition House 905.662.9040
D12 Optical Center 905.560.4174
D7 Shoppers Drug Mart 905.560.4900
E6 The Body Shop 905.561.6976
F7 Trade Secrets 905.561.7500

HOUSE & HOME
G8 Chelsea Home Furniture . 905.573.3408
A8 Govindo Galleries 905.560.9393
G11 HomeSense 905.561.2301
F8 Kitchen Stuff Plus 905.561.5151

JEWELLERY
C3 Charm Diamond Centres . 905.578.2882
A16 Fine Touch Jewelers 905.561.8242
A14 Mappins 905.561.4614
B10 Peoples 905.561.5740
K3 Vivah 905.561.7364

LEATHER & LUGGAGE
A13 Bentley 905.560.6572

MUSIC & VIDEOS
B8 HMV 905.560.3571

RESTAURANTS
Pad Jack Astor's 905.560.5254
F3 Sunset Grill 905.573.8688

SERVICES
K7 Island Ink Jet 905.578.4110
K16 London Gold Buyer . . . . . N/A
K13 Lotto Booth 905.573.8000
A1 Marlin Travel 905-561-5715
K1 My Lucky Lotto Shop 905.578.8000
D7 Post Office (inside Shoppers) . . . 905.560.4800 x4
C14 Quick Heel (shoe repair) 905.560.7300
G1 Stitch It 905-561-6273
D15 Things Remembered 905.561.6311

SPECIALTY FOODS
E11 Freshly Squeezed . . . . . N/A
G5 Laura Secord 905.578.5380
Pad LCBO 905.560.4970
Pad The Beer Store 905.560.9240

SPORTING GOODS
F1 SportChek 905.560.9371

Eastgate SQUARE
CLOSE TO HOME
75 Centennial Parkway North
905.561.2444 | eastgatesquare.ca

Cont. –in a resturant

Cont. in a library

Many other examples

- Collections in a Museum,
- Products in shops,
- Services in travel agency
- ….

Catalogue

Directory

Index

Inventory

portfolio

registry

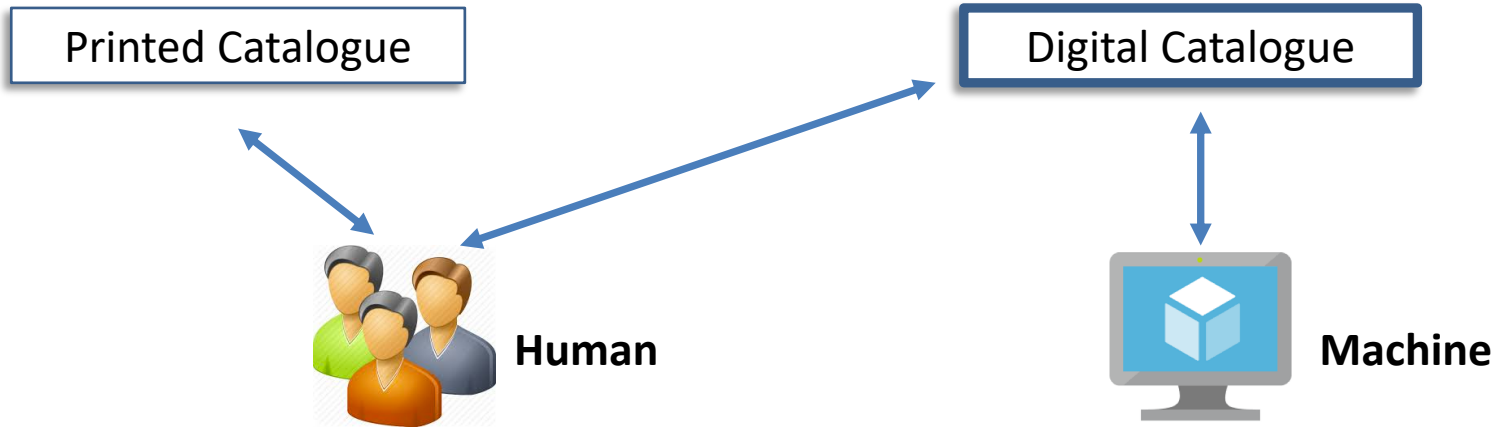
list

….

….

What is a catalogue (catalog)?

- A catalogue is **a list** of **things** such as the **goods** you can buy from a particular company, the **objects** in a museum, or the **books** in a library.

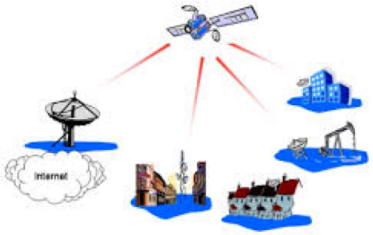


For machine: Catalogue as a service

- What is a service

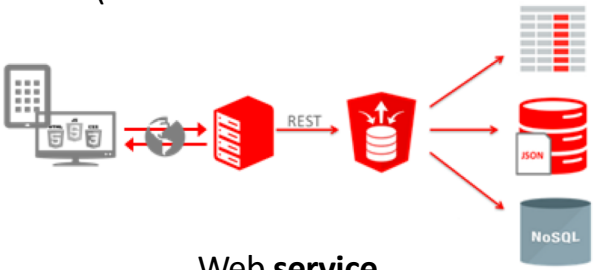
Holiday **services** (travel agency)

Logistics **service**



Satellite **service**

Communication **service**

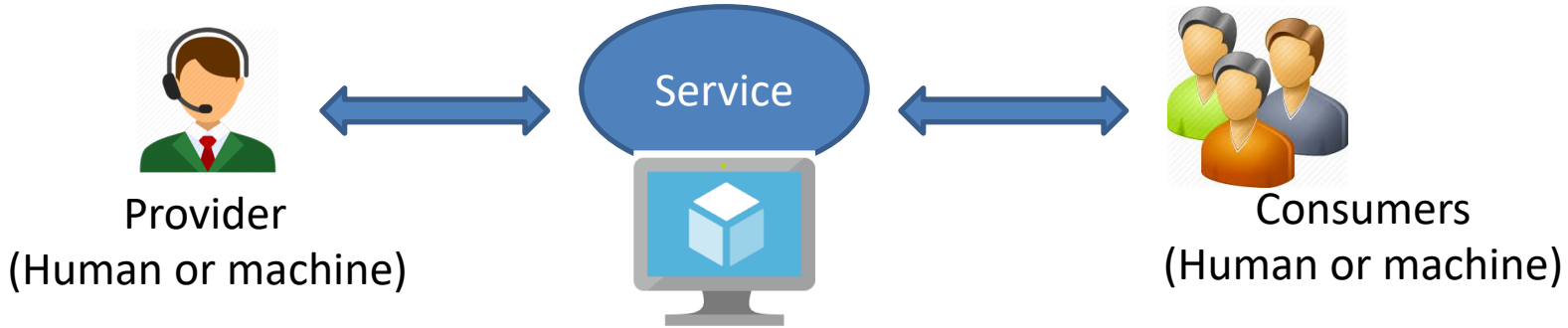


Web **service**

Cloud **service**

From business point of view

- **Services** (also known as "**intangible goods**") include attention, advice, access, experience, and affective labor. [*from wikipedia*]



Catalogue as a service

Service:

objective, function, interaction model, price etc.

Provider:

Consumer:

….

Discussion

- Can a catalogue provider and be different from the item providers?

Catalogue in data sciences: user stories

As a Researcher I want to **locate duplicates** of **a specific collection** compare them, to figure out possible **mislabeling**.

As A Researcher I want to know which **specimens** (species & region) **not yet sequenced** so that I can request material for DNA analysis and subsequent phylogeographic analyses.

As A Researcher I want to **choose specific specimens** from a list of specimens available **in my country** so that I can request loans or vi

As A Researcher I want to know **who** is in charge of **the collection** so that I can ask **specific questions**.

Do you have any user stories?

Catalogue for research assets

- Assets
- How to catalogue them?
  - Create?
  - Operate?
  - Use?

Assets- discussion

Assets in an ENVRI RI

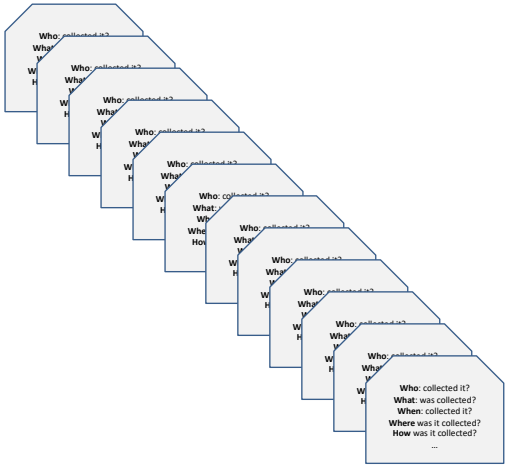
- Observation stations
- Devices
- Observations
- Measurements
- Models
- Software tools
- Services
- Workflows
- Training materials
- eInfrastructure resources
- Etc.

How to?
- simple example

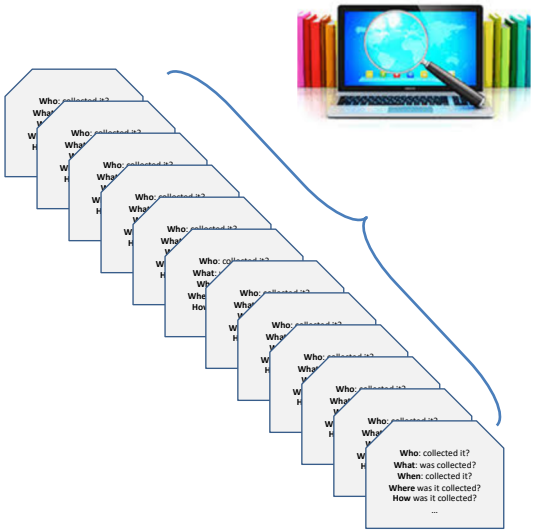
How to?
Step 1: create metadata information of items

Who: collected it?
What: was collected?
When: collected it?
Where was it collected?
How was it collected?
...

How to?
Step 2: organize the items



How to?
Step 3: provide interface for search



Discussion: requir

- Step one:

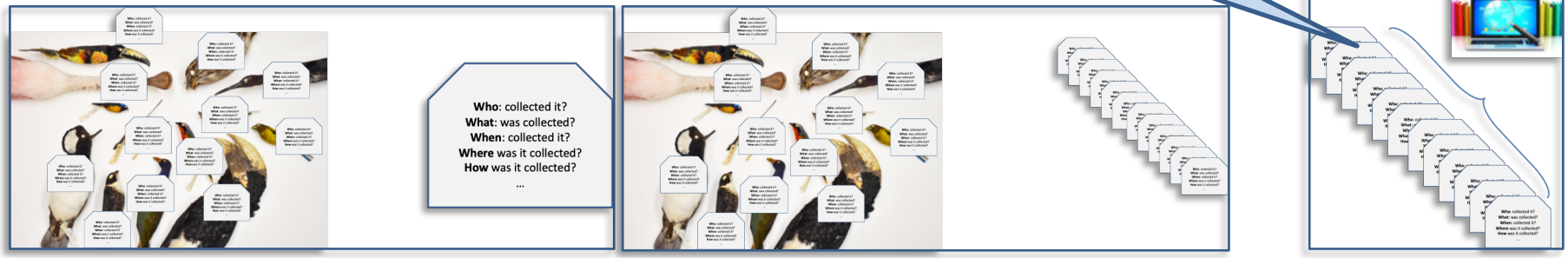
- **Metadata/meta information has to be *"Rich"***
- Metadata **schema (structure** of the metadata**)**
- Metadata **standards (standards of *metadata / metadata schema*)**

Discussion: requirements

- Step one:
- Step two:

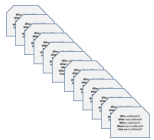
Discussion: requirements

- Step one:
- Step two:
- Step three:

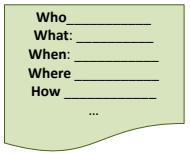


From information viewpoint: information flow in a digital catalogue

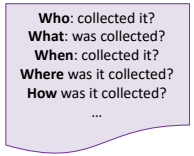
Printed catalogue



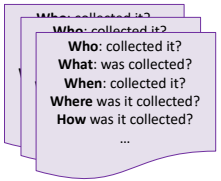
Digital catalogue



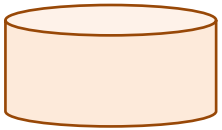
MetaData schema, standard



MetaData (record/ collection)



Registry (metadata)



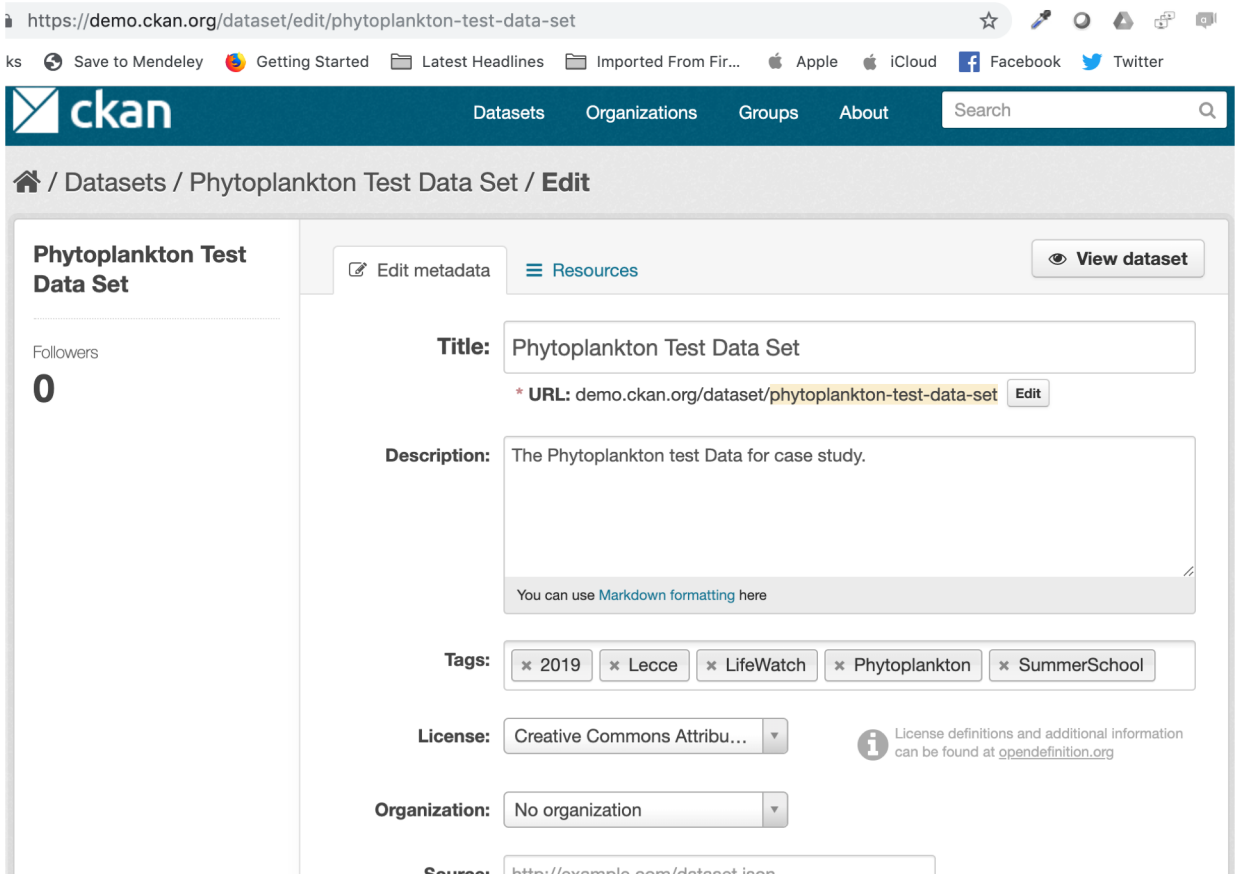
Repository (Assets)

Technologies

- The **Comprehensive Knowledge Archive Network (CKAN)** is
  - a web-based open-source management system for the storage and distribution of open data,
  - a powerful data catalogue system that is mainly used by public institutions seeking to share their data with the general public.
- Open source, Python web app, PostgreSQL DB, GPL
- https://demo.ckan.org/

Create data set

- Edit metadata
- Metadata
  - Title
  - Description
  - Tag
  - Organizations
  - License
  - Identifier



Metadata in CKAN

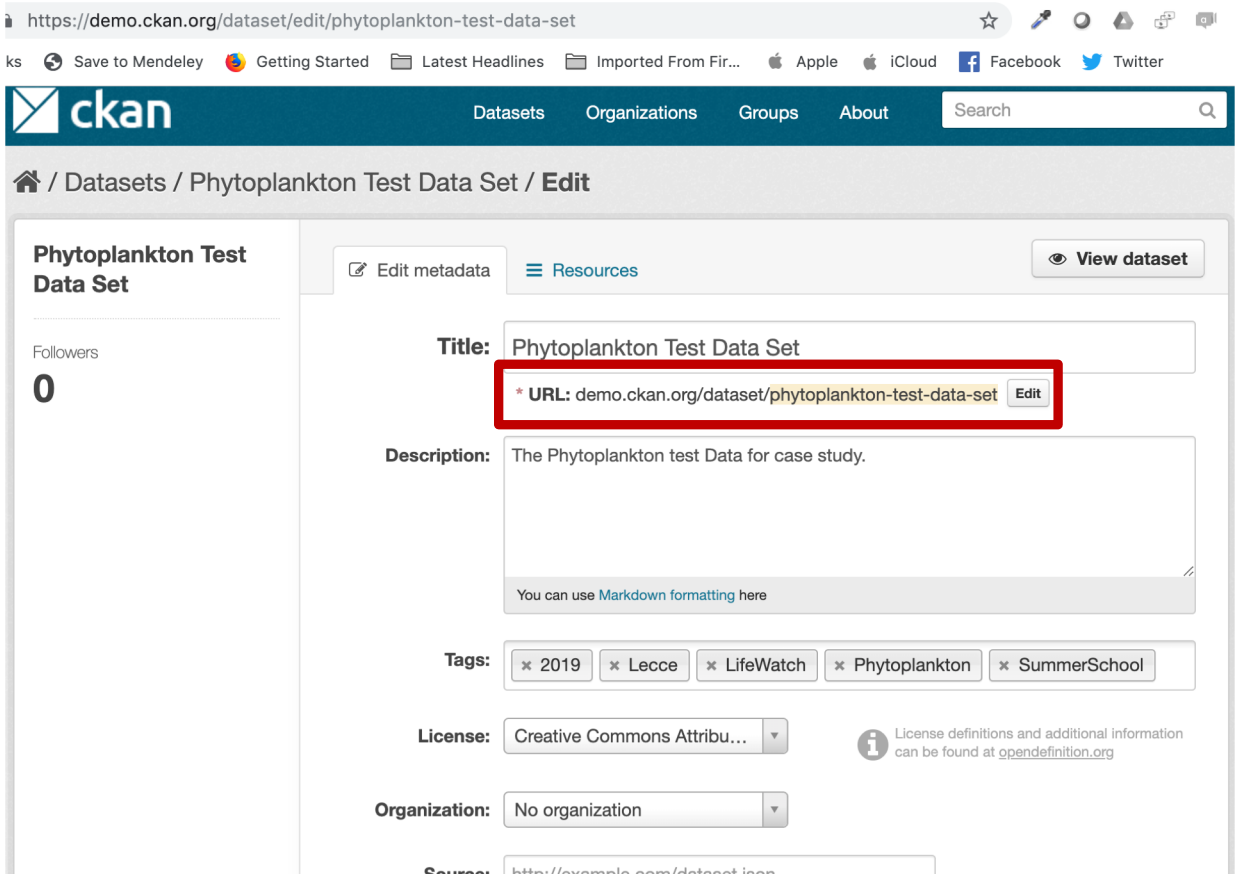
- **Title** – allows intuitive labelling of the dataset for search, sharing and linking.
- **Unique identifier** – dataset has a unique URL which is customizable by the publisher.
- **Groups** – display of which groups the dataset belongs to if applicable. Groups (such as science data) allow easier data linking, finding and sharing amongst interested publishers and users.
- **Description** – additional information describing or analysing the data. This can either be static or an editable wiki which anyone can contribute to instantly or via admin moderation.
- **Data preview** – preview .csv data quickly and easily in browser to see if this is the dataset you want.
- **Revision history** – CKAN allows you to display a revision history for datasets which are freely editable by users (as is thedatahub.org)
- **Extra fields** – these hold any additional information, such as location data (see geospatial feature) or types relevant to the publisher or dataset. How and where extra fields display is customizable.
- **Licence** – instant view of whether the data is available under an open licence or not. This makes it clear to users whether they have the rights to use, change and re-distribute the data.
- **Tags** – see what labels the dataset in question belongs to. Tags also allow for browsing between similarly tagged datasets in addition to enabling better discoverability through tag search and faceting by tags.
- **Multiple formats (if provided)** – see the different formats the data has been made available in quickly in a table, with any further information relating to specific files provided inline.
- **API key** – allows access every metadata field of the dataset and ability to change the data if you have the relevant permissions via API.

Other metadata standards related to catalogues

- Dublin CORE
- ISO 19115
- CKAN
- DCAT
- CERIF

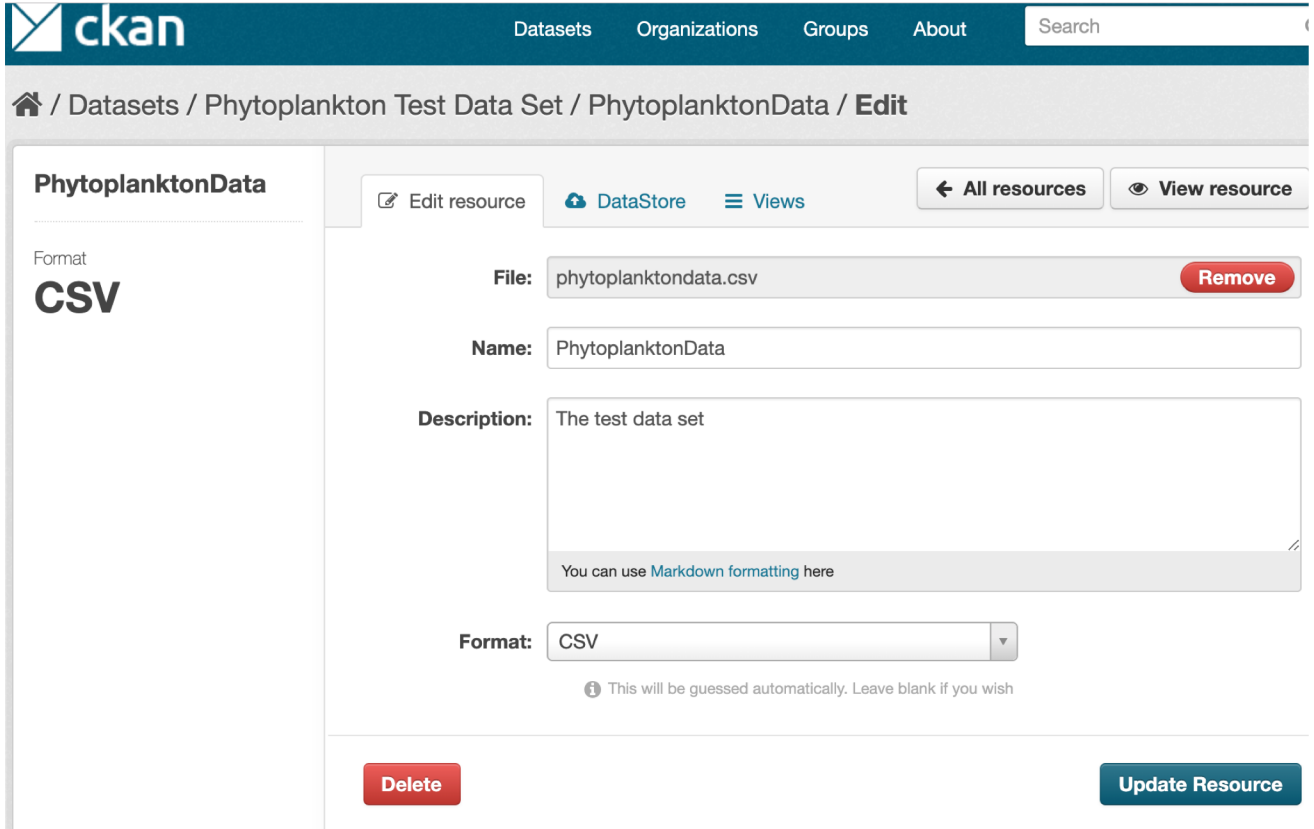
Create data set

- Edit metadata
- Metadata
  - Title
  - Description
  - Tag
  - Organizations
  - License
  - **Identifier**



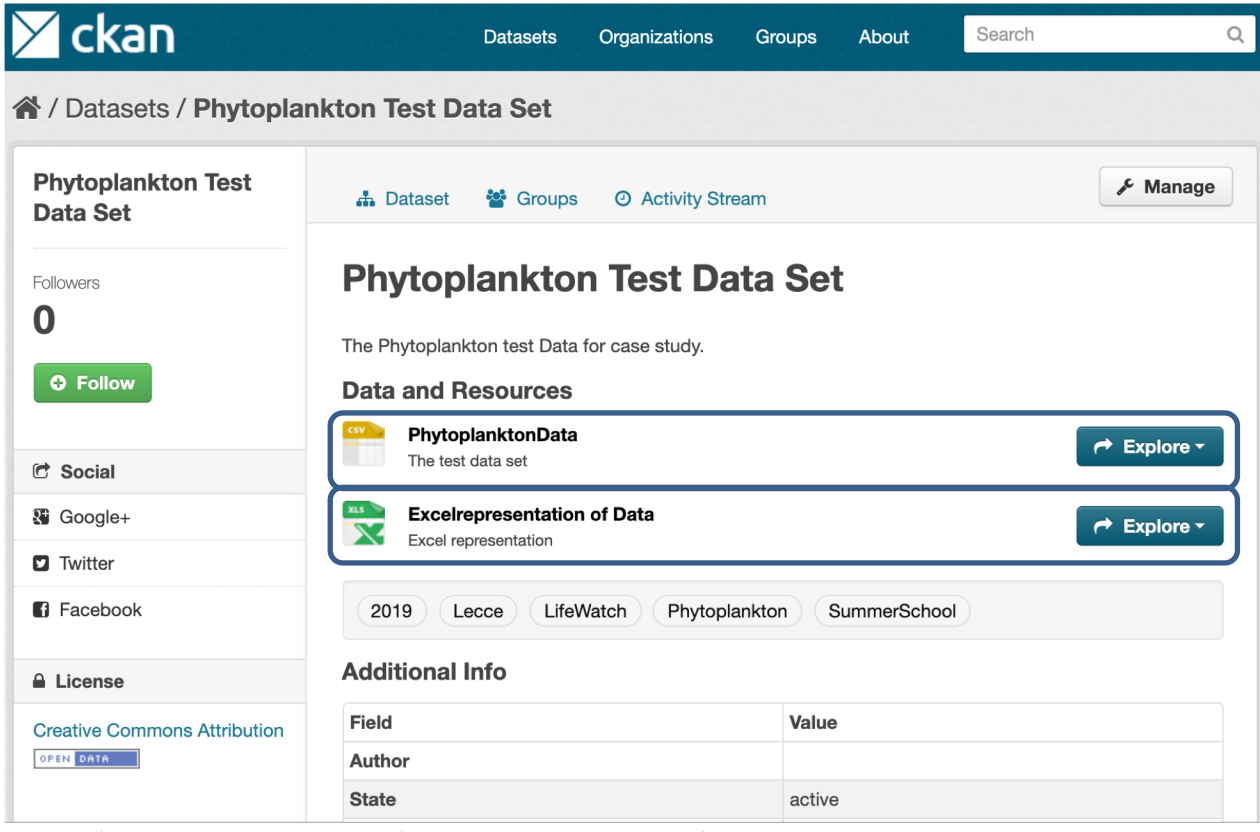
Create data set

- Edit metadata
- Resources
  - File
  - Name
  - Description
  - Format



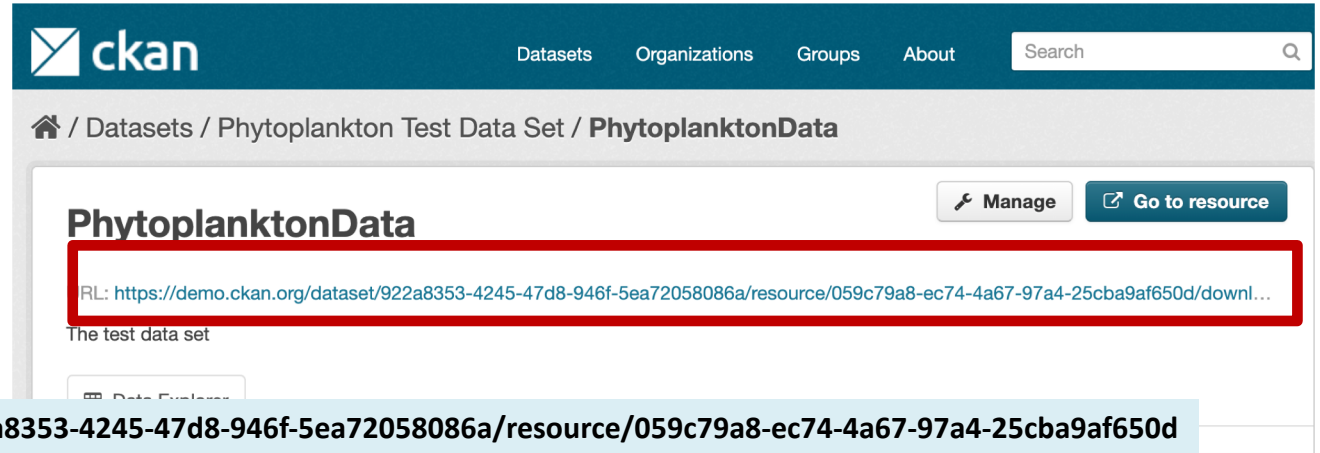
Data set contains multiple resources

- Edit metadata
- Resources
  - File
  - Name
  - Description
  - Format



Create data set

- Edit metadata
- Resources
  - Identifier



https://demo.ckan.org/dataset/922a8353-4245-47d8-946f-5ea72058086a/resource/059c79a8-ec74-4a67-97a4-25cba9af650d

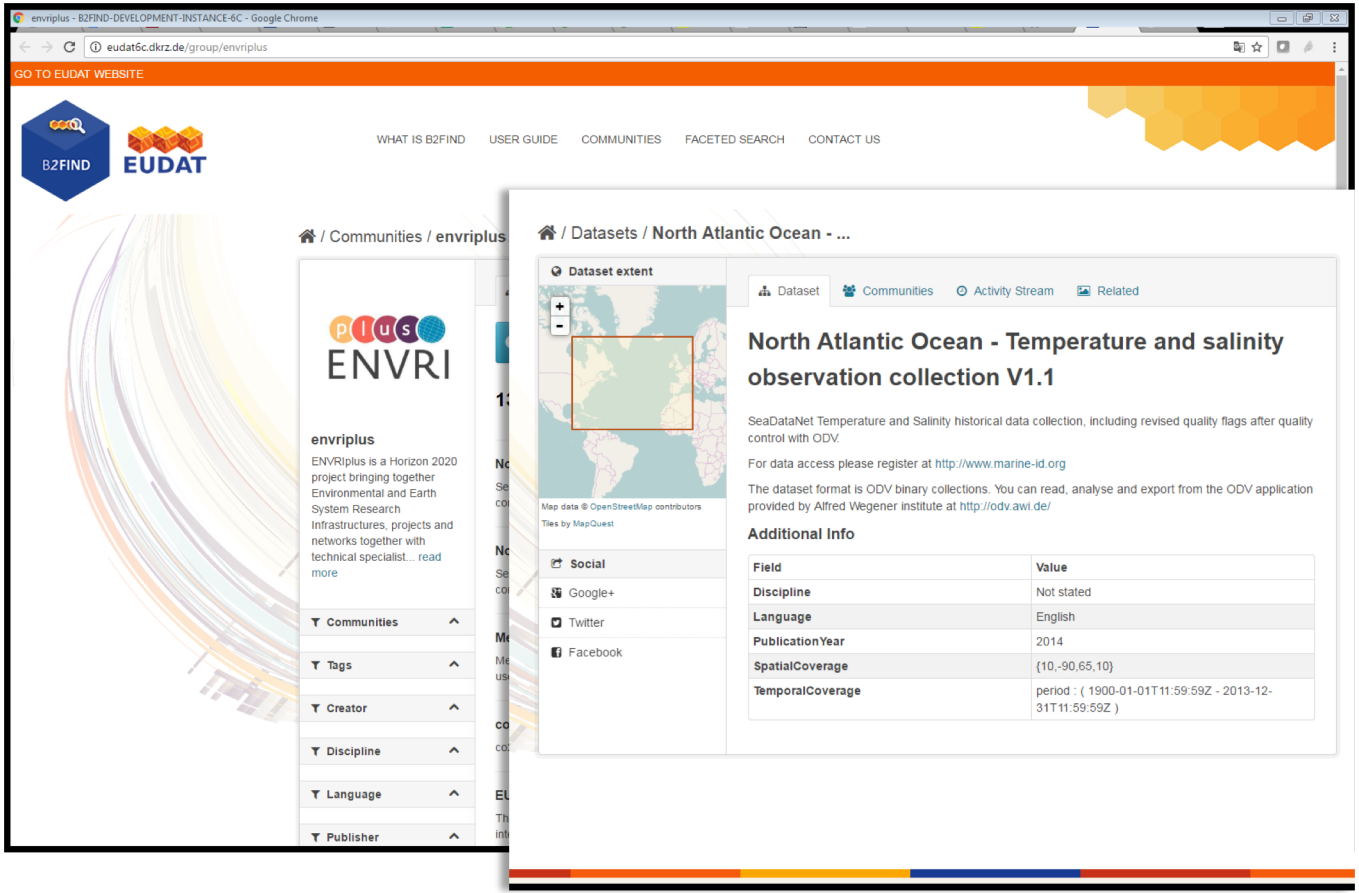
Metadata

Resource

the DEMONSTRATOR will HAVE:

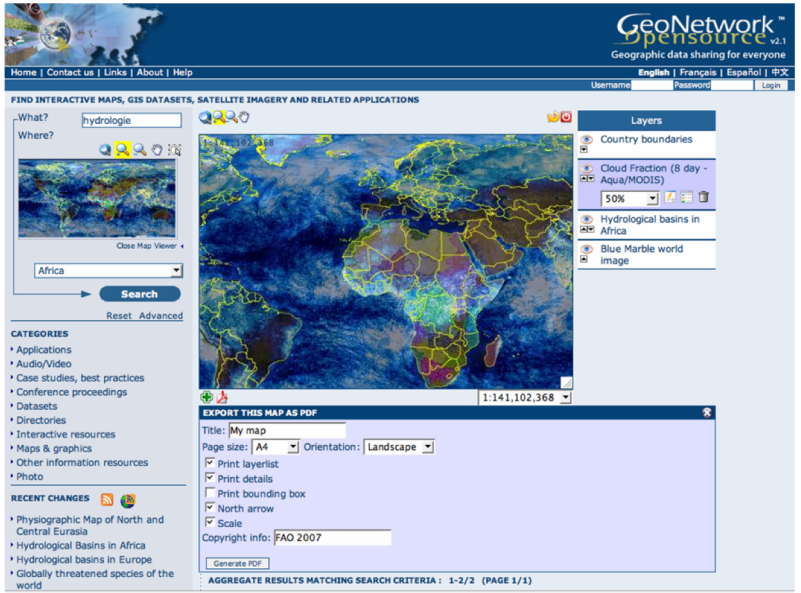
- Products from:
  - Euro-ARGO, SeaDataNet (IFREMER)
  - EPOS (NERC, INGV, GFZ)
  - ICOS (LU)
  - IAGOS (CNRS)
  - LTER (UBA-GmbH)
  - ANAEE (INRA)
  - EMBRC (MBA)
- Show them in EUDAT/B2FIND infrastructure implemented by DKRZ with CKAN software (up and running system, flexible, open-source and popular, quick win, and an RI-neutral service).

Achievements

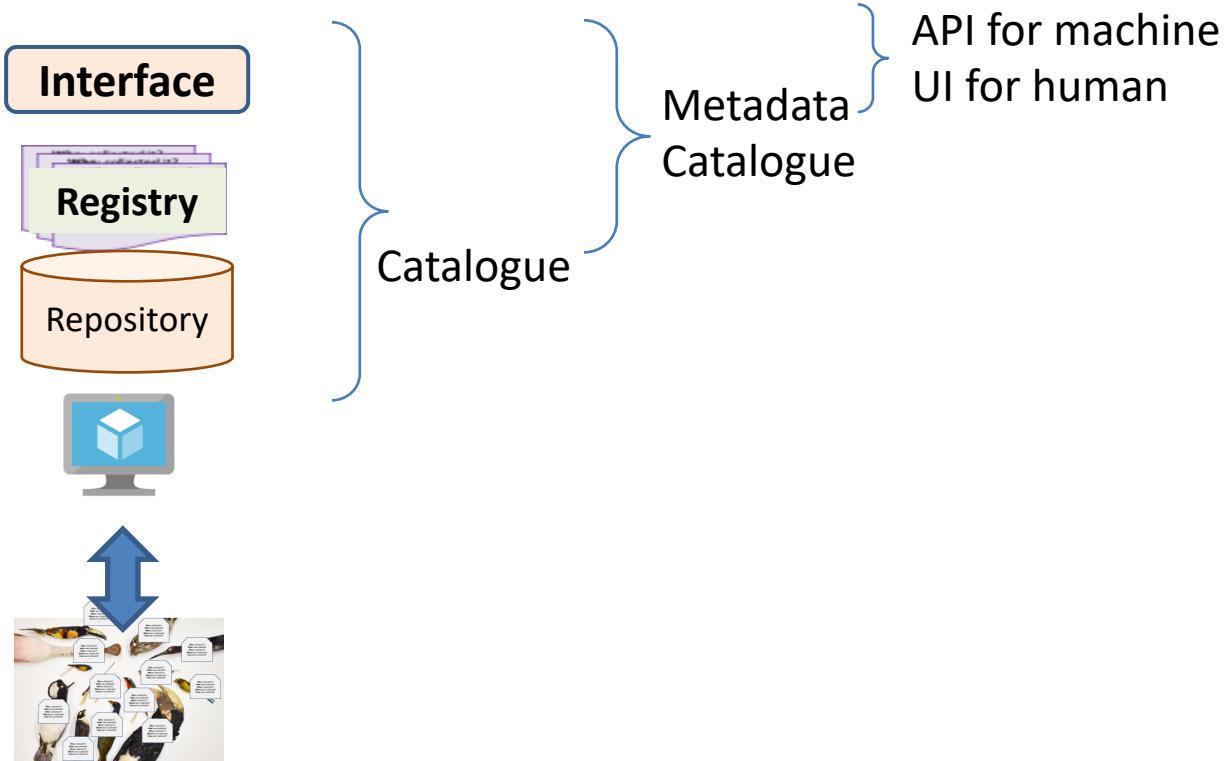


Technologies

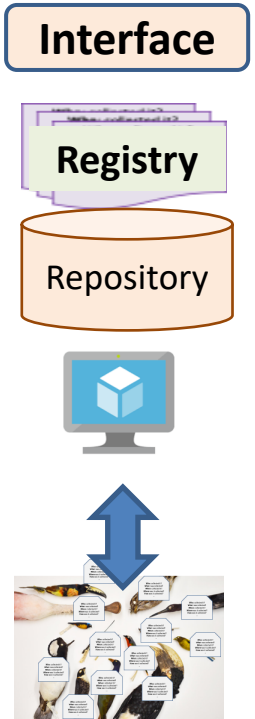
- The GeoNetwork project started out in year 2001 as a **Spatial Data Catalogue System** for the Food and Agriculture organisation of the United Nations (**FAO**), the United Nations World Food Programme (**WFP**) and the United Nations Environmental Programme (**UNEP**).
- At present the project is widely used as the basis of **Spatial Data Infrastructures** all around the world.
- The project is part of the **Open Source Geospatial Foundation (OSGeo)** and can be found at GeoNetwork opensource.



Catalogue in a research infrastructure

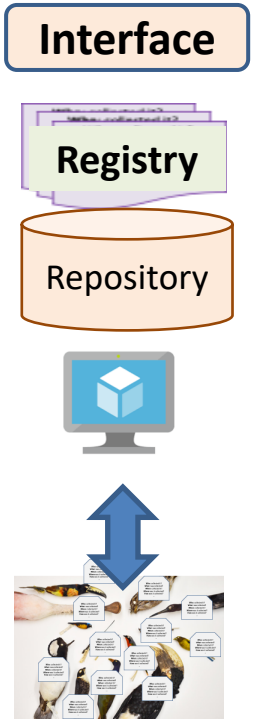


Cont.



Files, Relational data base, graph database, multi dimension arrays etc.

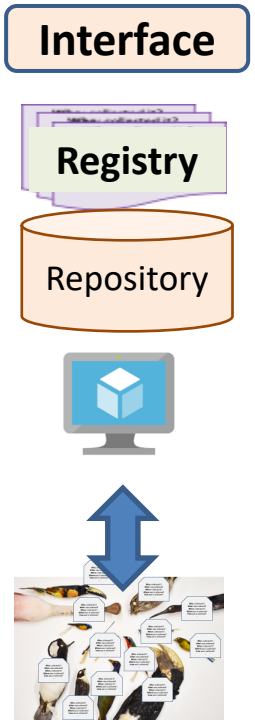
Approach 1: Metadata



Metadata (descriptive, e.g., Dublin core, CKAN, CERIF, observations, e.g., ISO19115, etc.),. **PID** (DOI, etc.) of the actual asset, **ID** of the record

Files, Relational data base, graph database, multi dimension arrays etc.

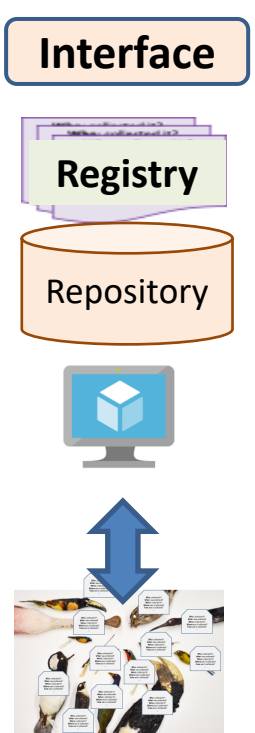
Approach 2: Linked open data, RDF



Linked open data and RDF:
Store records as RDF graph

Files, Relational data base, graph database, multi dimension arrays etc.

Cont.

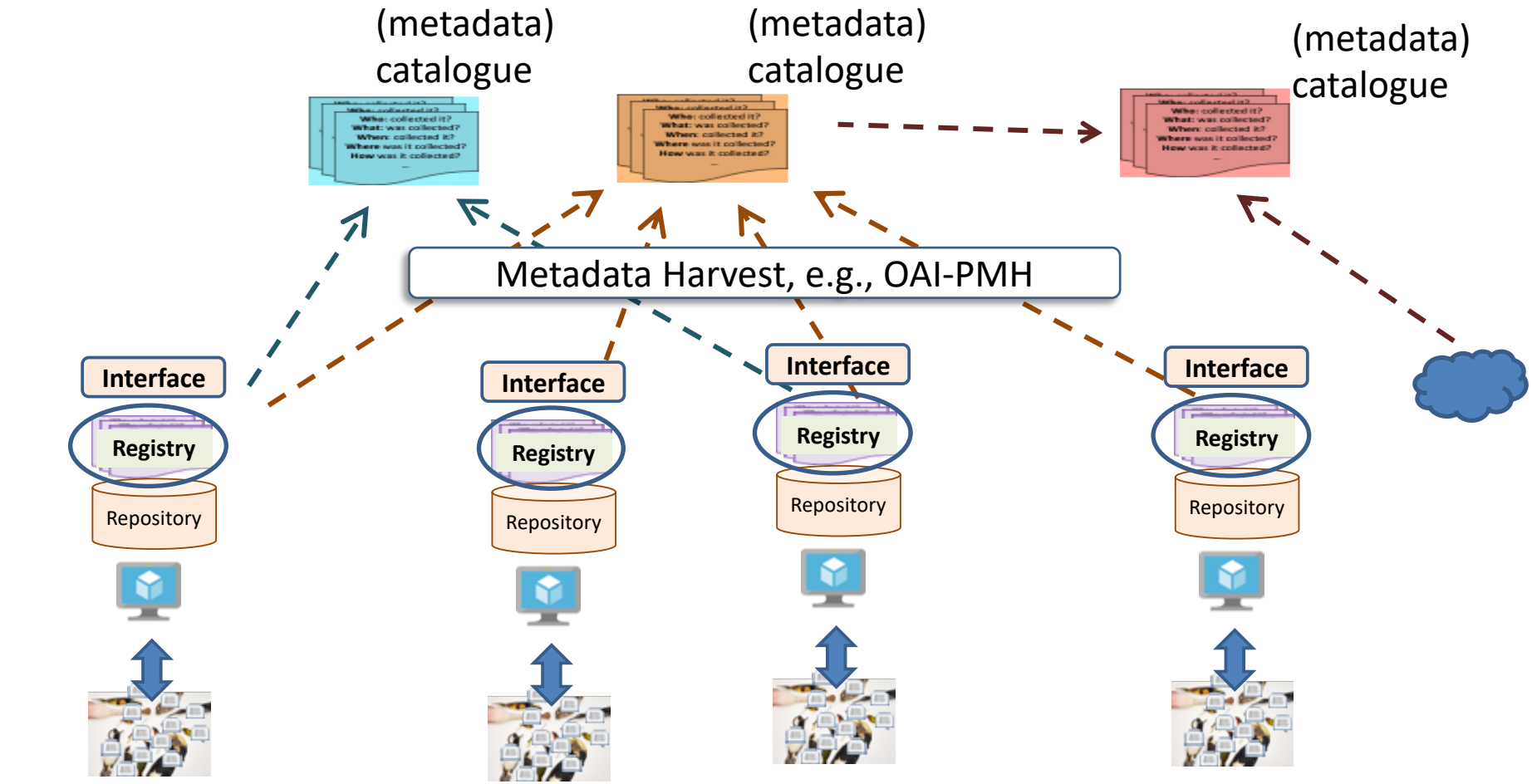


For administrator, for user, and for other services (catalogues, workflows, search engines, etc.)

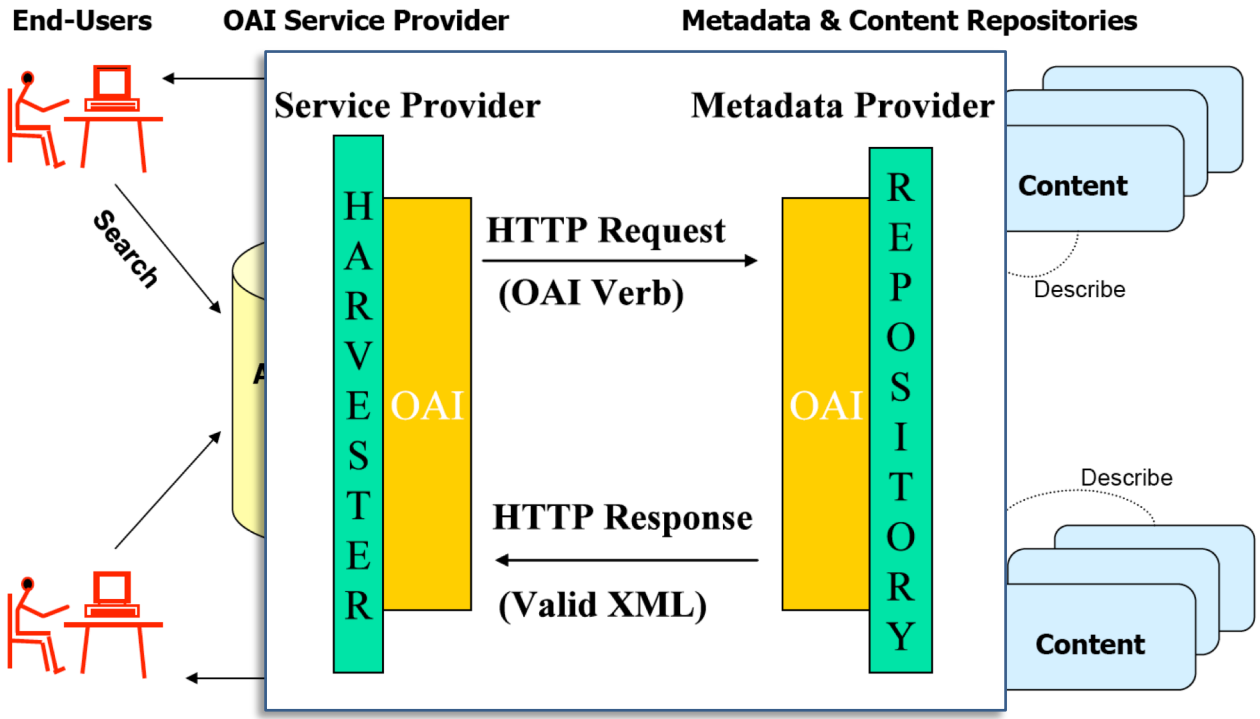
Discussion

- How are catalogue and PID related?

Catalogue for publishing and discovery



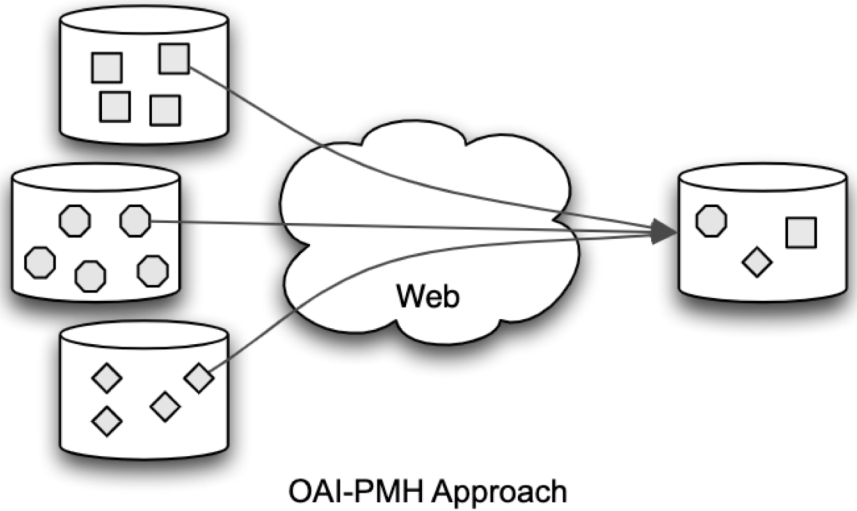
How does harvest work?



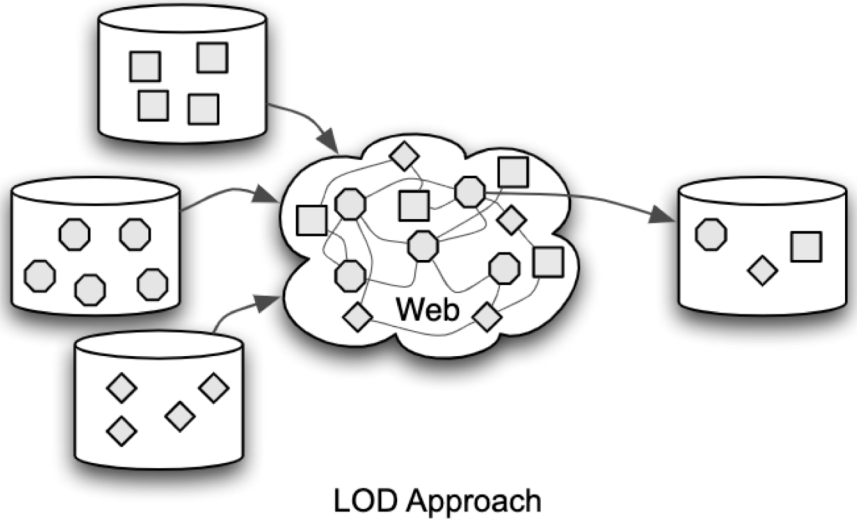
https://www.carli.illinois.edu/sites/files/digital_collections/documentation/OAIPMH_Cole_20100928.pdf

Other solutions

- Linked Open data approach



OAI-PMH Approach



LOD Approach

Discussion

- Does catalogue mean open access?

Discussion

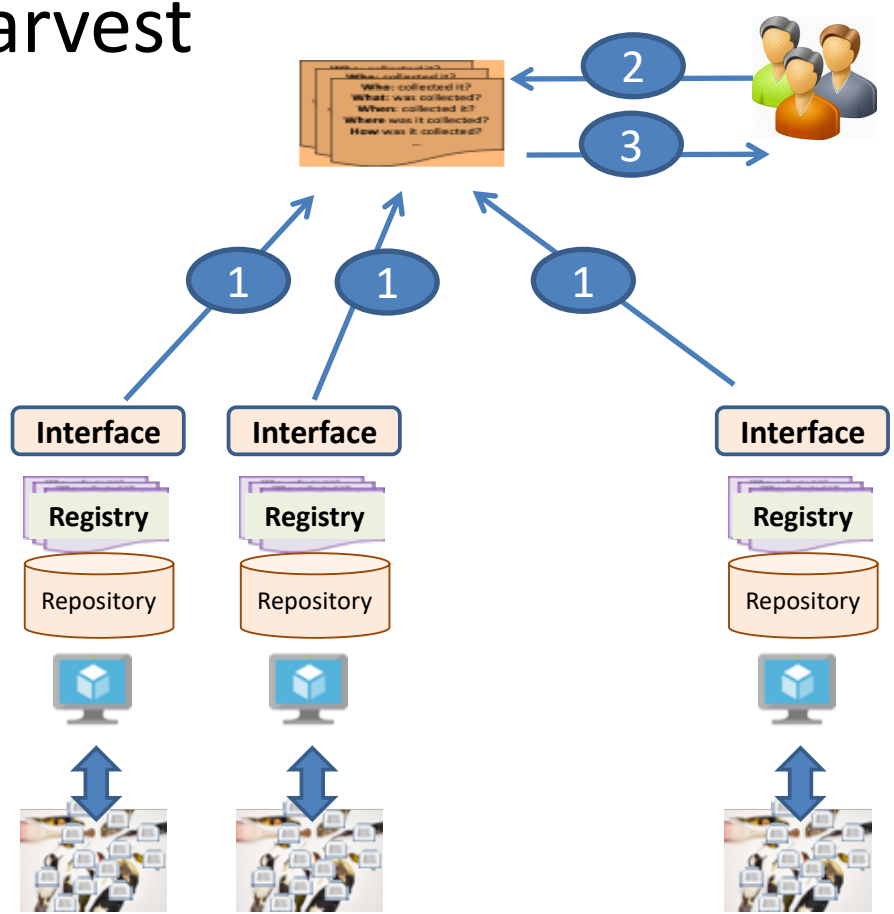
- Does catalogue mean open access?
- Catalogue is effective for：
  - Findability
  - Accessibility
  - Interoperability
  - Reusability

Discussion: how to search across catalogue

- Approaches:

Approach one: catalog harvest

- Harvest distributed catalogues
- Query the harvested catalogue



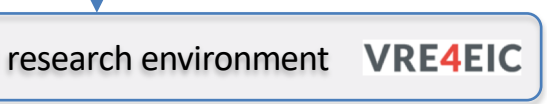
Cross RI RI using

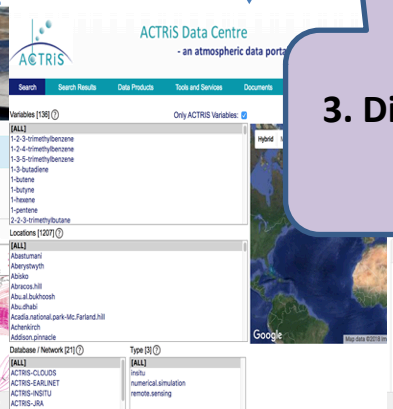
1. Map individual RI metadata onto CERIF **(Metadata Manager)**

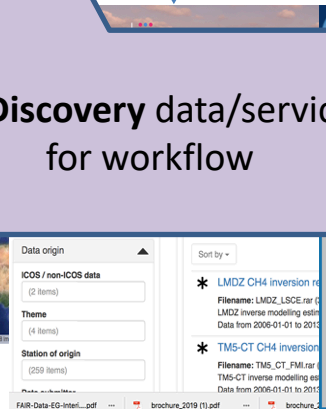
2. Import CERIF records into a eVRE based catalogue (**Interoperability manager**)

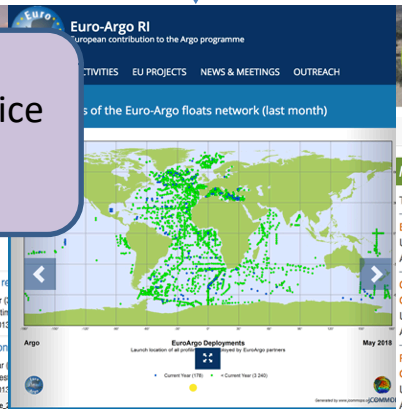
3. Discovery data/service for workflow

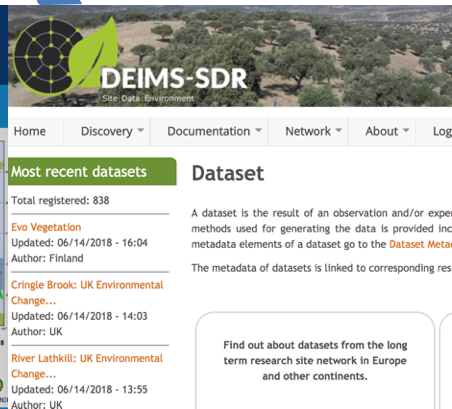
4. Execute workflow and manage scientific experiments (**Workflow manager/etc.**)



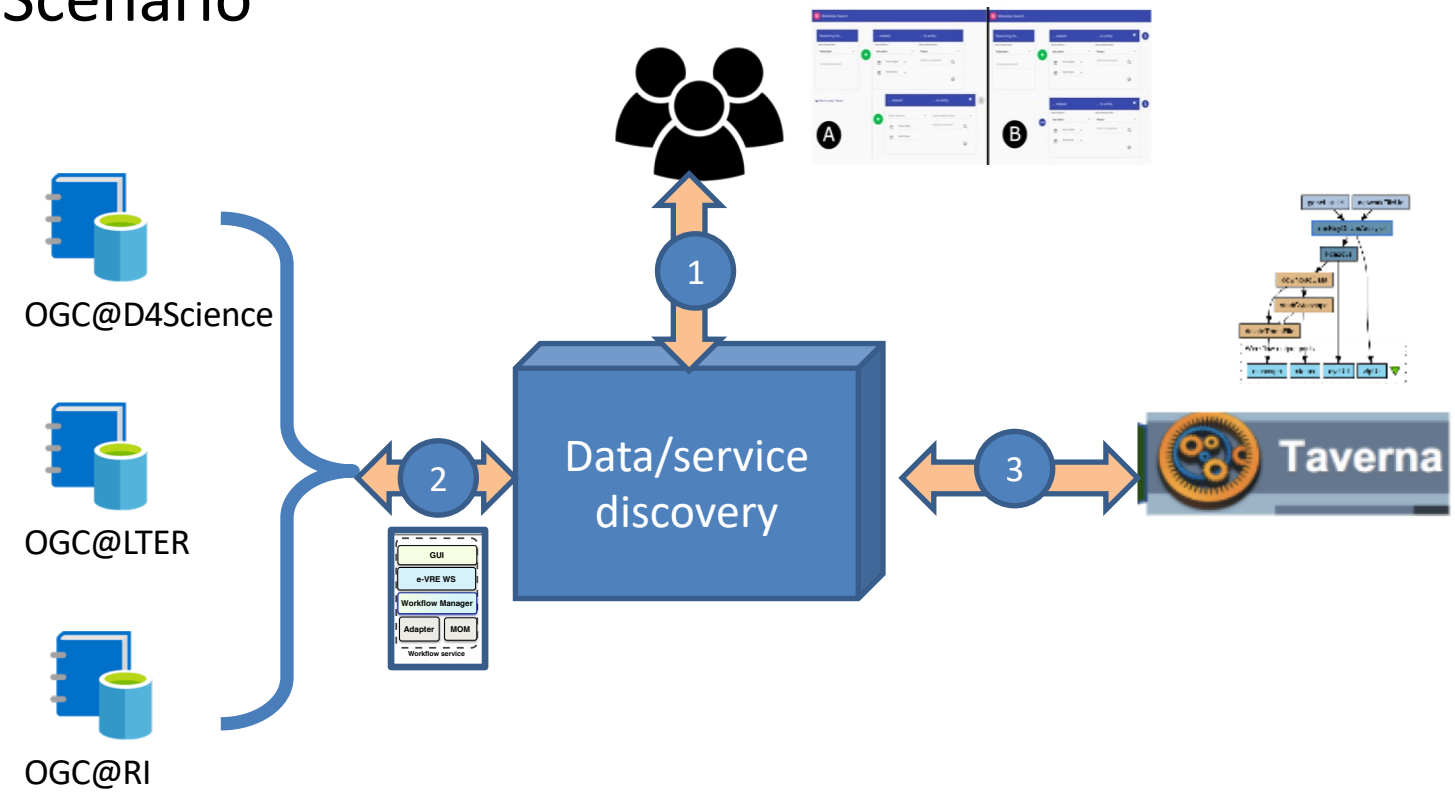






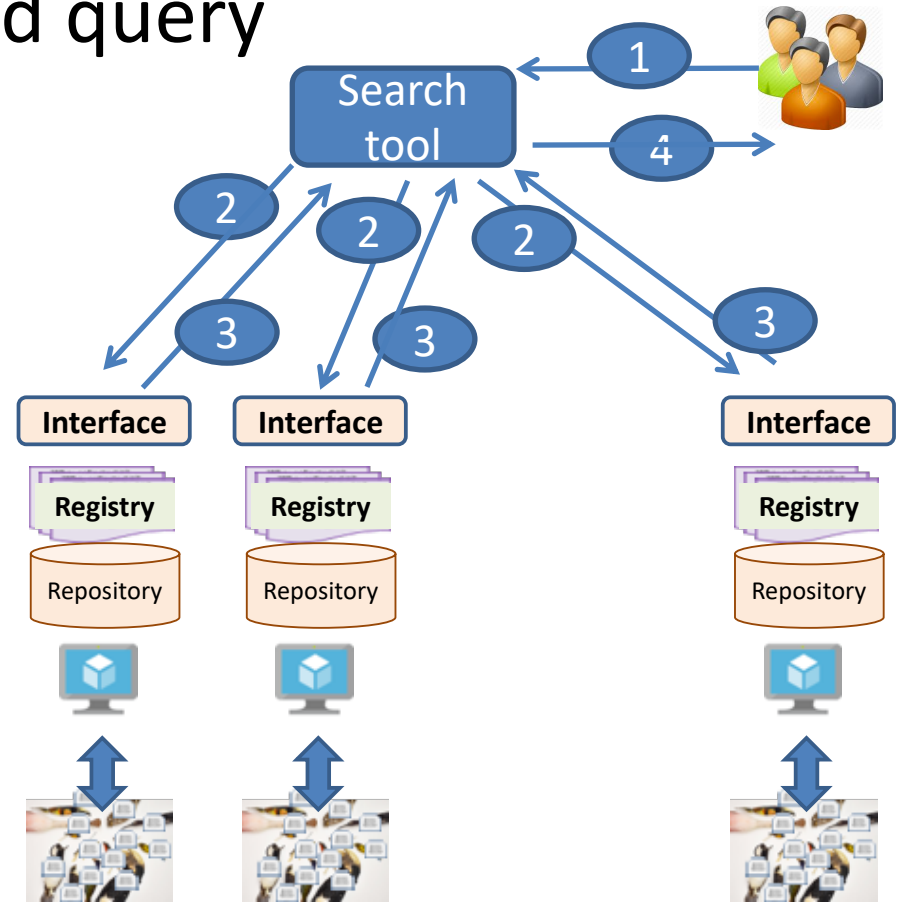


Scenario


OGC@D4Science
OGC@LTER
OGC@RI
1
2
3
Data/service discovery
GUI
e-VRE WS
Workflow Manager
Adapter
MOM
Workflow service
A
B
Taverna

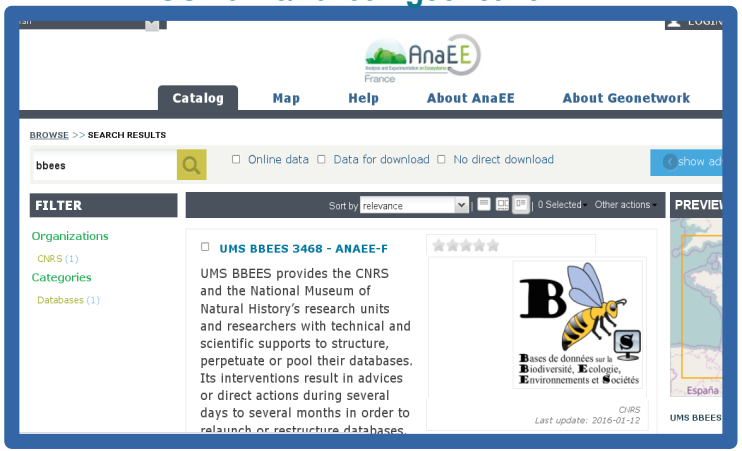
Approach two: distributed query

- Generate queries to distributed catalogues
- Collect the results

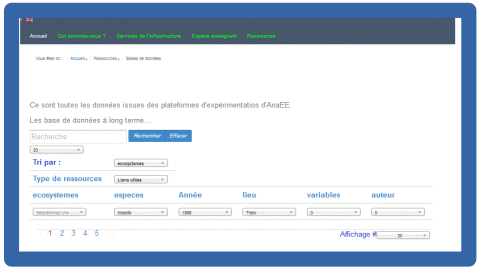


Discovery catalogue

ISO 19115/19139 - geonetwork



Dedicated portal for access to resources

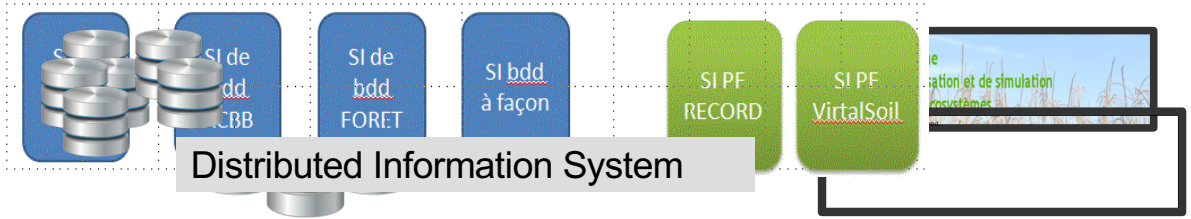


Exp. serv.

Keywords **thesaurus**

Annotation **ontology**

Identification citation



Distributed Information System

Exercise (20 minutes)

- Objective: learn the basic interface of CKAN
- Step -1
  - Go to https://demo.ckan.org/
  - Register an account
  - Login in the account

Exercise (20 minutes)

- Step -2
  - Create an organization, create a group

Exercise (20 minutes)

- Step -3
  - Create a data set (metadata)
  - Upload data
  - Add data set to the group

Exercise (20 minutes)

- Step -4
- API
  - Query: https://demo.ckan.org/api/search/dataset?q=course
  - Replace “course” with your own keywords
  - Try to find your own data set

Exercise (20 minutes)

- Step -5
- API
  - Get: https://demo.ckan.org/api/rest/dataset/testdataset0630
  - Get your own data set via the interface

Summary

- Further reading material
  - Erickson, J., Maali, F.: Data catalog vocabulary (DCAT). W3C recommendation, W3C, http://www.w3.org/TR/2014/ REC-vocab-dcat-20140116/ (2014).
  - DataCite 2017 - DATACITE METADATA WORKING GROUP: DataCite Metadata Schema Documentation for the Publication of Research Data. Online available: 10.5438/0014 (2017).
  - Haslhofer, B., Klas, W.: A survey of techniques for achieving metadata interoperability, ACM Computing Surveys (CSUR), vol. 42, no 2, http://eprints.cs.univie.ac.at/79/1/haslhofer08_acmSur_final.pdf (2010)
  - Hodgson, C.: N800R1, Where to start - advice on creating a metadata schema or application profile, TC46 SC 11 Interest Group / N Documents https://groups.niso.org/apps/group_public/download.php/7272/N800R1%20Where%20to%20start-advice%20on%20creating%20a%20metadata%20schema.pdf (2011)
- Continue with the CKAN experiments