

DATA FAIRNESS

LECCE
1-5 JUL
2019

INTERNATIONAL SUMMER SCHOOL
FOR ENVIRONMENTAL & EARTH SCIENCE
INFRASTRUCTURES
LIFEWATCH.EU/ISS-DATA-FAIRNESS



Catalogue and cataloguing

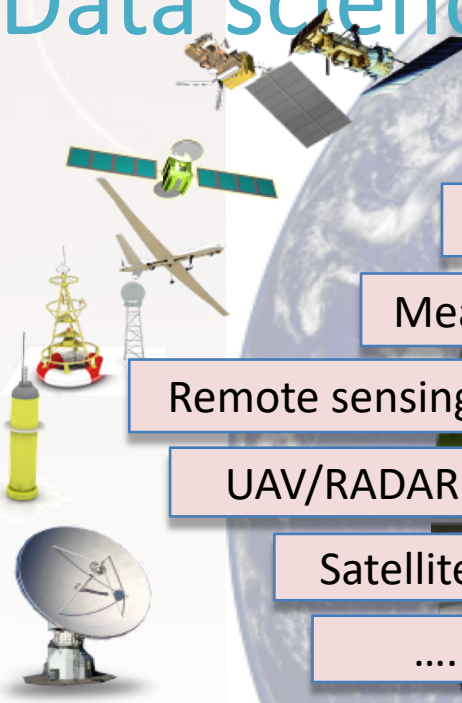
Zhiming Zhao

z.zhao@uva.nl



UNIVERSITY OF AMSTERDAM

Data sciences in environmental sciences



What?...

How?...

$$\begin{aligned}
 \rho \frac{Dv_x}{Dt} &= -\frac{\partial p}{\partial x} + \mu \Delta v_x + f_x \\
 \rho \frac{Dv_y}{Dt} &= -\frac{\partial p}{\partial y} + \mu \Delta v_y + f_y \\
 \rho \frac{Dv_z}{Dt} &= -\frac{\partial p}{\partial z} + \mu \Delta v_z + f_z
 \end{aligned}$$

Observe

invers models

Atmospheric

Marine

Complex system modelling

Measure

Model composition

Remote sensing

Multi scale modelling

UAV/RADAR



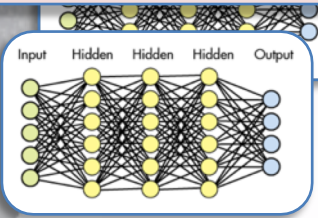
Data centric modelling..

Satellite

....

Biodiversity
ecosystems

Mid earth



cloud

What-if?...

Cloud/Edge/Fog

Workflow

HTC/HPC

BigData / AI

...





Cannot find

An effective catalogue is needed!

Outline

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

What is a catalogue?

- in a mall

DIRECTORY OF SHOPS & SERVICES

General Avenue
Kestrel Avenue
Centennial Parkway
Lower Level

ATM
Elevator
Entrance
Food Court
Lockers
Nursing Room
Public Washrooms
Stairs
Telephone

MALL HOURS
Monday to Friday 10:00am to 8:00pm
Saturdays 9:30am to 6:00pm
Sundays 11:00am to 6:00pm

CUSTOMER SERVICE CENTRE
To reach Customer Service, call 905.561.2444 open all shopping hours. Customer Service is located between Precipita Jewellers and Charm Diamond Centres.

GIFT CARDS
Available for purchase at the Customer Service Centre. We accept Cash, VISA, MasterCard & Interac. There are no processing fees or additional charges when purchasing the card. Redeemable at most stores and services and valid at Eastgate Square only.

EASTGATE SQUARE MANAGEMENT OFFICES
The Eastgate Square Management Offices are located on the second floor, up the stairs located near the public washrooms.

February 2015

[Facebook](#)
[Twitter](#)
[Instagram](#)

| | | | | |
|--|------------------|---|--------------------------|-------------|
| APPEL - ATHLETIC WEAR | 05.678.5063 | 05.688.5301 | Ward Mobile | 05.688.5301 |
| A11 Lids | 05.561.1111 | 05.661.0242 | Wintersweat | 05.661.0242 |
| F1 Sportchek | 05.560.9271 | 05.578.0800 | C13 Wireless Valley | 05.578.0800 |
| | | 05.560.5758 | K15 WOV! mobile boutique | 05.560.5758 |
| APPEL - CHILDREN'S | 05.573.3908 | FASHION ACCESSORIES | | |
| C8 The Children's Place | 05.573.3908 | 05.602.2592 | | |
| APPEL - LADIES | | FAST FOOD | | |
| A5 Alo n' Tally | 05.561.1022 | A3 Andy | 05.920.7999 | |
| 08 Anna Bella | 05.560.1288 | 02 Jeremy the Cook | 05.061.1000 | |
| B1 Armani | 05.963.0509 x182 | 02 Kang Poo Wok | 05.578.8882 | |
| 03 Chelsea | 05.561.4028 | 06 New York Fried | 05.578.5200 | |
| E6 Edipak | 05.578.1470 | D14 Orange Julius/DQ | 05.560.7421 | |
| F4 Gorge | 05.560.6944 | A2 Pizza Pops | 05.560.9104 | |
| 04 La Senza | 05.560.0180 | D1 Subway | 05.573.8636 | |
| E1 Le Chateau | 05.673.8980 | D5 The Express | 05.560.8788 | |
| E9 Northern Reflections | 05.573.7155 | D3 Tim Hortons | 05.561.8534 | |
| C7 Bellman's | 05.560.4810 | K10 Tim Hortons Kiosk | 05.560.8444 | |
| 07 Sury Shair | 05.561.4203 | FINANCIAL SERVICES | | |
| A16 Zoytek | 05.578.4238 | B12 C.I.B.C. Banking Centre | 05.561.3600 | |
| APPEL - MEN'S | | D16 Continental Currency Exchange | 05.561.6100 | |
| E19 International Clothing | 05.573.7480 | 05 Liberty Tax | 05.560.6990 | |
| C16 Moores Clothing for Men | 05.561.4265 | FOOTWEAR | | |
| APPEL - UNISEX | | 07 Foot Locker | 05.578.5063 | |
| 09 Blanches | 05.560.4408 | 04 Payless Shoe Source | 05.682.0832 | |
| 05 Fuschacher/Schacherm | 089.965.8374 | C14 Quik Hill | 05.560.7300 | |
| A8 Book Zone | 05.297.2056 | C4 SoftMac | 05.560.7483 | |
| FS Urban Planet | 05.578.2677 | 04 Steps | 05.930.8978 | |
| BOOKS & NEWS | | GENERAL MERCHANDISE & VARIETY | | |
| E3 Coles | 05.560.3784 | Dalton's | 05.578.1740 | |
| D10 International News | 05.561.1841 | A4 Showco | 05.578.5222 | |
| K17 Smokers Dan | 05.882.0928 | GROCERY STORES | | |
| CARDS, STATIONERY & GIFTS | | 05.561.5291 | | |
| A7 Carlton Cards | 05.882.0217 | HEALTHYWEAR & PERSONAL CARE SERVICES | | |
| B2 Sponsor Gifts | 05.560.4863 | K9 Beauty 21 Shoe Studio | NA | |
| DEPARTMENT STORES | | D11 Dr. Jang (OPTOMETRIST) | 05.561.1110 | |
| F2 Winners | 05.545.4741 | C1 Eastgate Dental Centre | 05.560.2714 | |
| ELECTRONICS, PHOTOGRAPHY & PHONES | | A10 CMC | 05.561.5324 | |
| 016 Bell | 05.578.2140 | C13 Good Expectations | 05.573.0770 | |
| D11 ES Games | 05.682.8410 | 011 HairCutters | 05.573.8952 | |
| K12 Kiko | 05.560.8160 | C13 Hearing Solutions | 05.560.8585 | |
| K2 In Touch | 05.797.8452 | 02 L'Amour Nails | 05.578.6245 | |
| A12 Knoods | 05.560.6720 | 03 Le Salon & Spa | 05.560.2900 | |
| E7 Rogers | 05.560.7027 | C10 LensCutters | 05.578.3292 | |
| D17 Tooth | 05.584.1206 | D15 MasterCuts | 05.882.7222 | |
| D9 Value | 05.560.7339 | D18 Mutation Hair | 05.882.8040 | |
| A8 The Source | 05.560.0289 | D12 Optical Centre | 05.560.4174 | |
| K14 Virgin Mobile | 05.573.7039 | D7 Shoppers Drug Mart | 05.560.6900 | |
| | | 11 The Body Shop | 05.561.6976 | |
| | | F7 Trade Secrets | 05.561.7500 | |

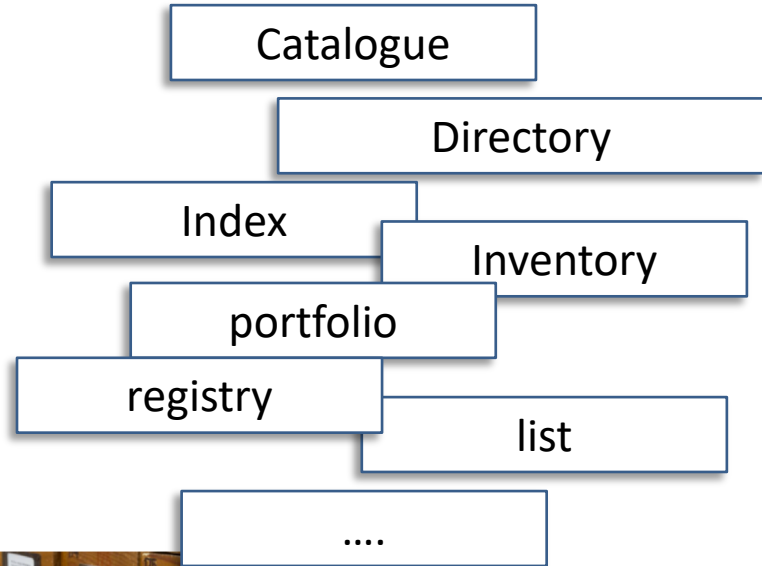


Cont. in a library



Many other examples

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

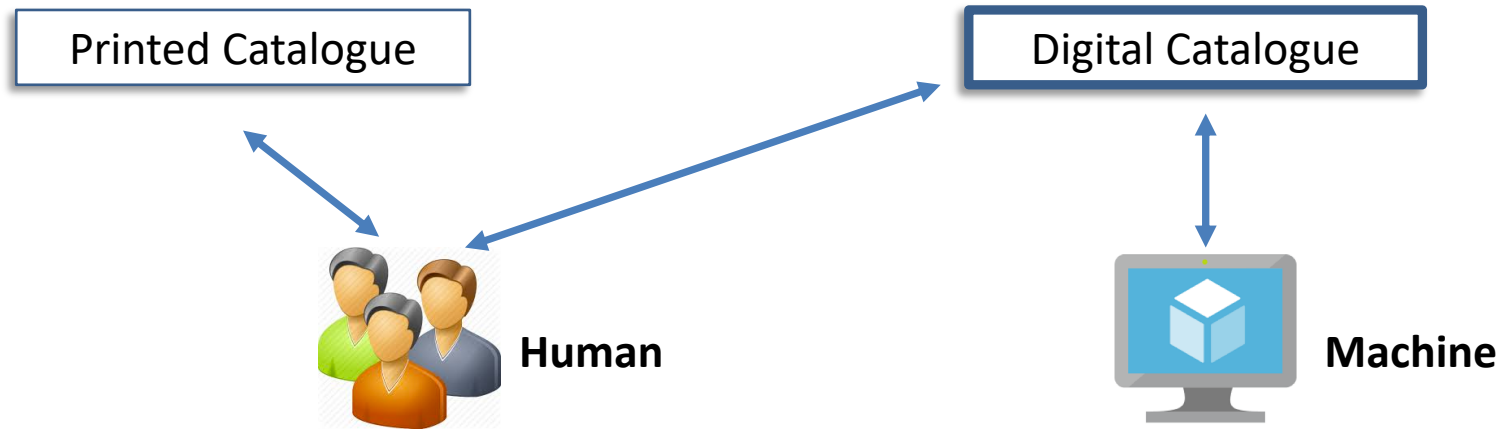


...

...

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)



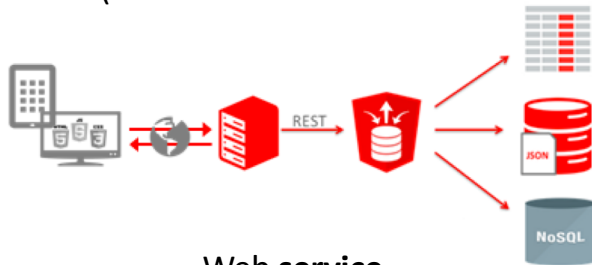
Logistics **service**



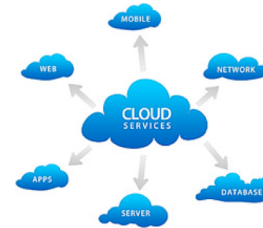
Satellite **service**



Communication **service**



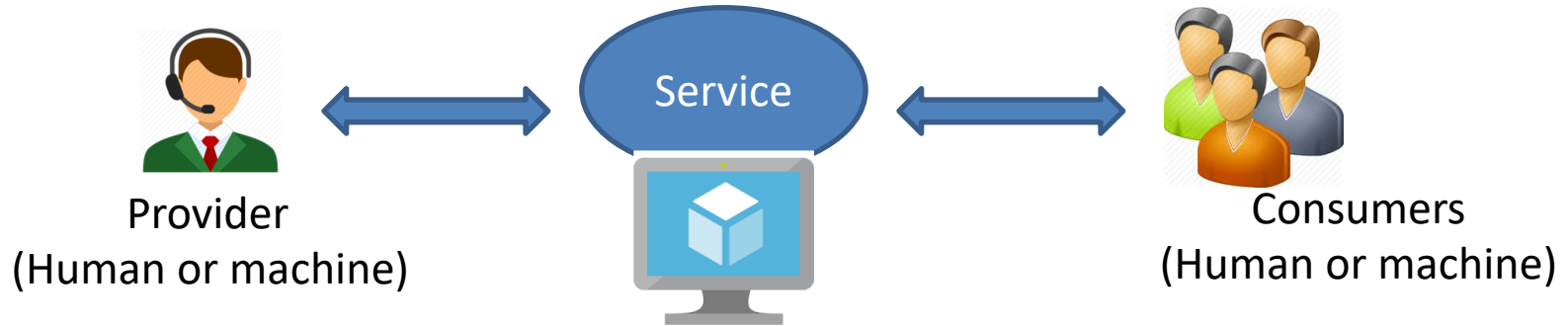
Web **service**



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:

....

Catalogue in data sciences: user stories

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

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 visit.

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.

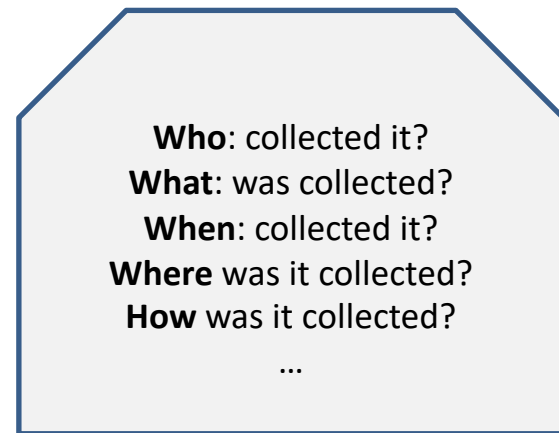
How to?

- simple example



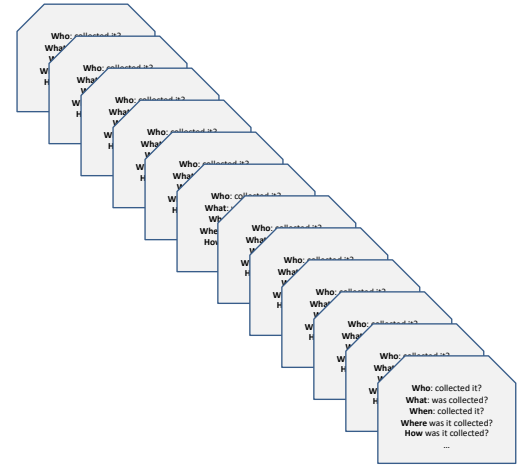
How to?

Step 1: create metadata information of items



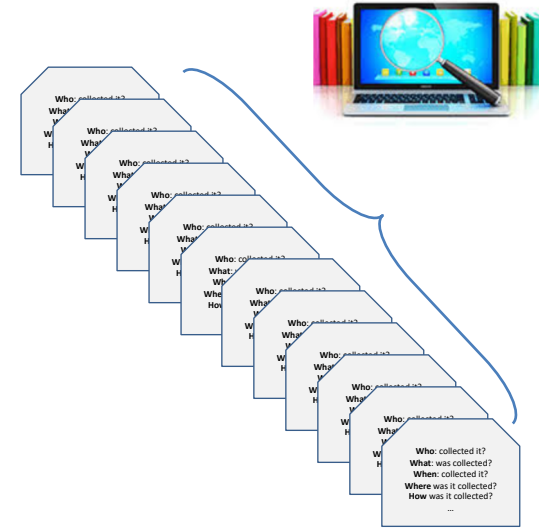
How to?

Step 2: organize the items



How to?

Step 3: provide interface for search



Discussion: requirements

- Step one:

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



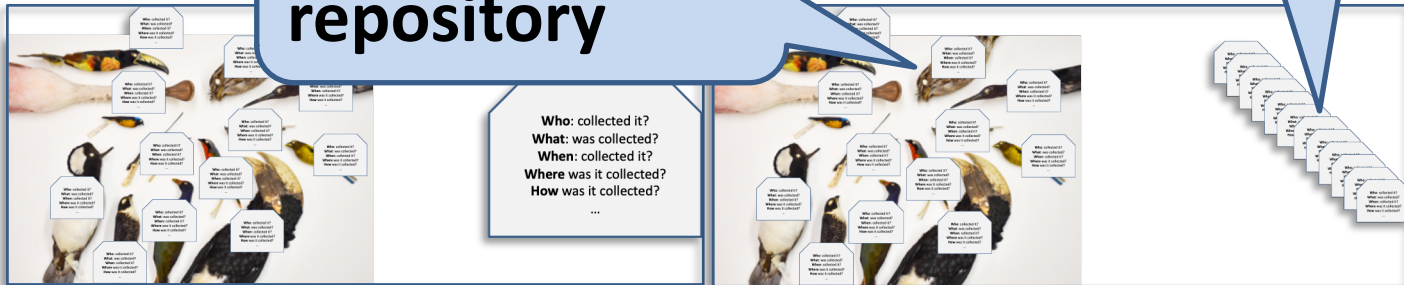
Discussion: requirements

- Step one:
- Step two:

**Data
repository**

**Metadata
registry**

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



Discussion: requirements

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

**Indexing,
Search**

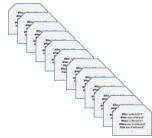


From information viewpoint: information flow in a digital catalogue

Printed catalogue

Who _____
What: _____
When: _____
Where _____
How _____
...

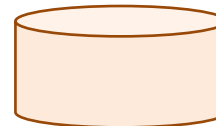
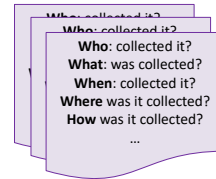
Who: collected it?
What: was collected?
When: collected it?
Where was it collected?
How was it collected?
...



Digital catalogue

Who _____
What: _____
When: _____
Where _____
How _____
...

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



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

The screenshot shows the CKAN 'Edit dataset' page for 'Phytoplankton Test Data Set'. The browser address bar shows the URL: <https://demo.ckan.org/dataset/edit/phytoplankton-test-data-set>. The CKAN logo is in the top left, and navigation links for 'Datasets', 'Organizations', 'Groups', and 'About' are in the top right. A search bar is also present. The breadcrumb trail is: Home / Datasets / Phytoplankton Test Data Set / Edit. On the left sidebar, the dataset title 'Phytoplankton Test Data Set' is displayed, along with a 'Followers' count of 0. The main content area has two tabs: 'Edit metadata' (active) and 'Resources'. A 'View dataset' button is in the top right of the main area. The 'Title' field contains 'Phytoplankton Test Data Set' and has an 'Edit' button next to it. Below the title, the 'URL' is shown as 'demo.ckan.org/dataset/phytoplankton-test-data-set' with an 'Edit' button. The 'Description' field contains 'The Phytoplankton test Data for case study.' and has a note: 'You can use Markdown formatting here'. The 'Tags' field shows five tags: '2019', 'Lecce', 'LifeWatch', 'Phytoplankton', and 'SummerSchool'. The 'License' field is set to 'Creative Commons Attribution...' with an information icon and a link to 'opendefinition.org'. The 'Organization' field is set to 'No organization'. The 'Source' field is partially visible at the bottom, showing 'http://example.com/dataset.json'.

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

The screenshot shows the CKAN 'Edit metadata' interface for a dataset. The browser address bar displays the URL `https://demo.ckan.org/dataset/edit/phytoplankton-test-data-set`. The CKAN logo and navigation menu are visible at the top. The dataset title is 'Phytoplankton Test Data Set'. The 'Title' field contains the same text. Below it, the 'URL' field is highlighted with a red box and contains the text `* URL: demo.ckan.org/dataset/phytoplankton-test-data-set` with an 'Edit' button. The 'Description' field contains the text 'The Phytoplankton test Data for case study.' The 'Tags' field shows five tags: '2019', 'Lecce', 'LifeWatch', 'Phytoplankton', and 'SummerSchool'. The 'License' field is set to 'Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International'. The 'Organization' field is set to 'No organization'.

Create data set

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

The screenshot shows the CKAN interface for editing a resource. The breadcrumb trail is: Home / Datasets / Phytoplankton Test Data Set / PhytoplanktonData / Edit. The resource name is 'PhytoplanktonData' and its format is 'CSV'. The description is 'The test data set'. The file name is 'phytoplanktondata.csv'. There are buttons for 'Edit resource', 'DataStore', 'Views', 'All resources', and 'View resource'. At the bottom, there are 'Delete' and 'Update Resource' buttons.

ckan Datasets Organizations Groups About Search

Home / Datasets / Phytoplankton Test Data Set / PhytoplanktonData / Edit

PhytoplanktonData

Format
CSV

Edit resource DataStore Views All resources View resource

File: phytoplanktondata.csv Remove

Name: PhytoplanktonData

Description: The test data set

You can use [Markdown formatting](#) here

Format: CSV

This will be guessed automatically. Leave blank if you wish

Delete Update Resource

Data set contains multiple resources

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

The screenshot shows the CKAN interface for the 'Phytoplankton Test Data Set'. The page includes a header with the CKAN logo and navigation links for Datasets, Organizations, Groups, and About. A search bar is located in the top right corner. The main content area features the dataset title, a description, and a 'Data and Resources' section. Two resources are listed: 'PhytoplanktonData' (CSV) and 'Excelrepresentation of Data' (Excel), both with 'Explore' buttons. The page also includes social media links, a license, and a table of additional information.

Phytoplankton Test Data Set

The Phytoplankton test Data for case study.

Data and Resources

- PhytoplanktonData** (CSV) - The test data set - Explore
- Excelrepresentation of Data** (XLS) - Excel representation - Explore

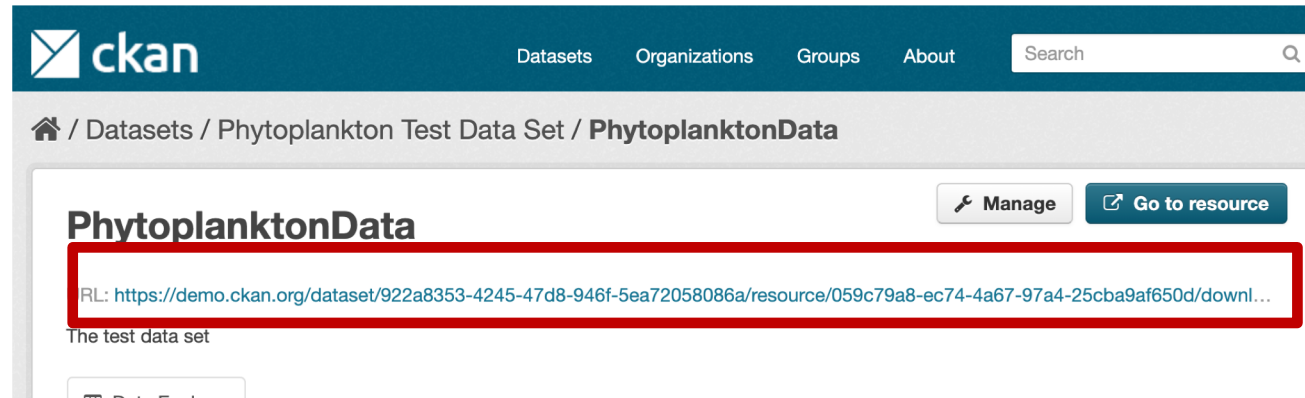
2019 Lecce LifeWatch Phytoplankton SummerSchool

Additional Info

| Field | Value |
|--------|--------|
| Author | |
| State | active |

Create data set

- Edit metadata
- Resources
 - Identifier



The screenshot shows the CKAN interface for a dataset named 'Phytoplankton Test Data Set'. The specific resource is 'PhytoplanktonData'. The URL for this resource is highlighted with a red box: <https://demo.ckan.org/dataset/922a8353-4245-47d8-946f-5ea72058086a/resource/059c79a8-ec74-4a67-97a4-25cba9af650d/download...>. The resource is described as 'The test data set'. There are 'Manage' and 'Go to resource' buttons visible.

<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

The screenshot displays the ENVIplus website interface. At the top, there is a navigation bar with the BZFIND and EUDAT logos, and a menu with items: WHAT IS BZFIND, USER GUIDE, COMMUNITIES, FACETED SEARCH, and CONTACT US. Below the navigation bar, the main content area is divided into two sections. On the left, there is a sidebar for the 'envriplus' community, featuring the 'PLUS ENVRI' logo and a description: 'ENVIplus is a Horizon 2020 project bringing together Environmental and Earth System Research Infrastructures, projects and networks together with technical specialist... read more'. Below the description are expandable filters for Communities, Tags, Creator, Discipline, Language, and Publisher. On the right, the main content area shows the 'Datasets / North Atlantic Ocean - ...' page. This page includes a 'Dataset extent' section with a map of the North Atlantic Ocean, a 'Social' section with links to Google+ and Facebook, and an 'Additional Info' section with a table of metadata.

envriplus - BZFIND-DEVELOPMENT-INSTANCE-6C - Google Chrome
eudat5c.dkrz.de/group/envriplus
GO TO EUDAT WEBSITE

WHAT IS BZFIND USER GUIDE COMMUNITIES FACETED SEARCH CONTACT US

GO TO EUDAT WEBSITE

envriplus
ENVIplus is a Horizon 2020 project bringing together Environmental and Earth System Research Infrastructures, projects and networks together with technical specialist... read more

Communities Tags Creator Discipline Language Publisher

North Atlantic Ocean - Temperature and salinity observation collection V1.1

Dataset extent

Map data © OpenStreetMap contributors
Tiles by MapQuest

Social
Google+
Twitter
Facebook

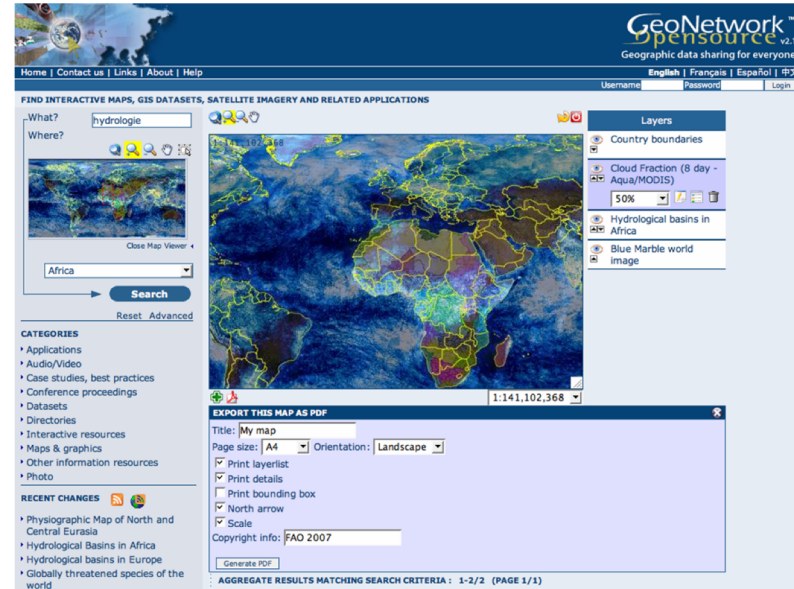
Additional Info

| Field | Value |
|------------------|--|
| Discipline | Not stated |
| Language | English |
| PublicationYear | 2014 |
| SpatialCoverage | {10,-90,65,10} |
| TemporalCoverage | period : (1900-01-01T11:59:59Z - 2013-12-31T11:59:59Z) |

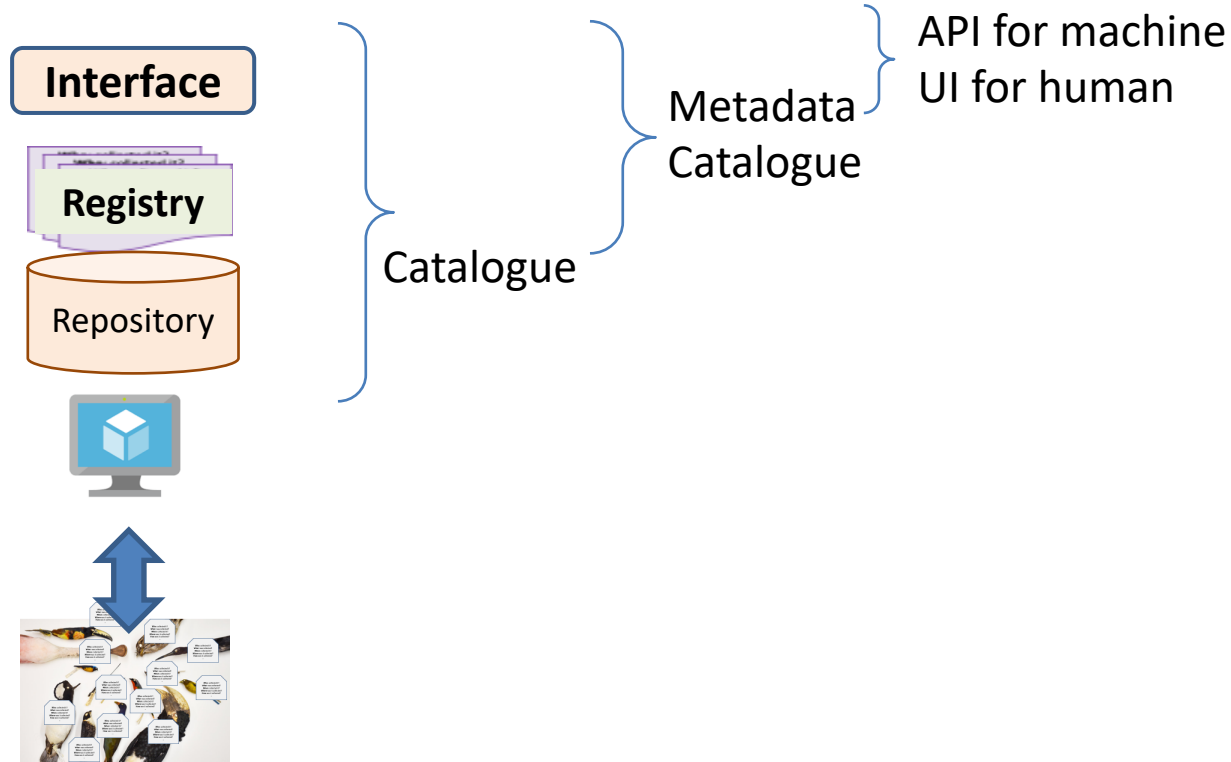
Technologies



- 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.

Interface

Registry

Repository

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



Approach 1: Metadata

Interface

Registry

Repository



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

Interface

Registry

Repository



**Linked open data and RDF:
Store records as RDF graph**

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

Cont.

Interface

Registry

Repository

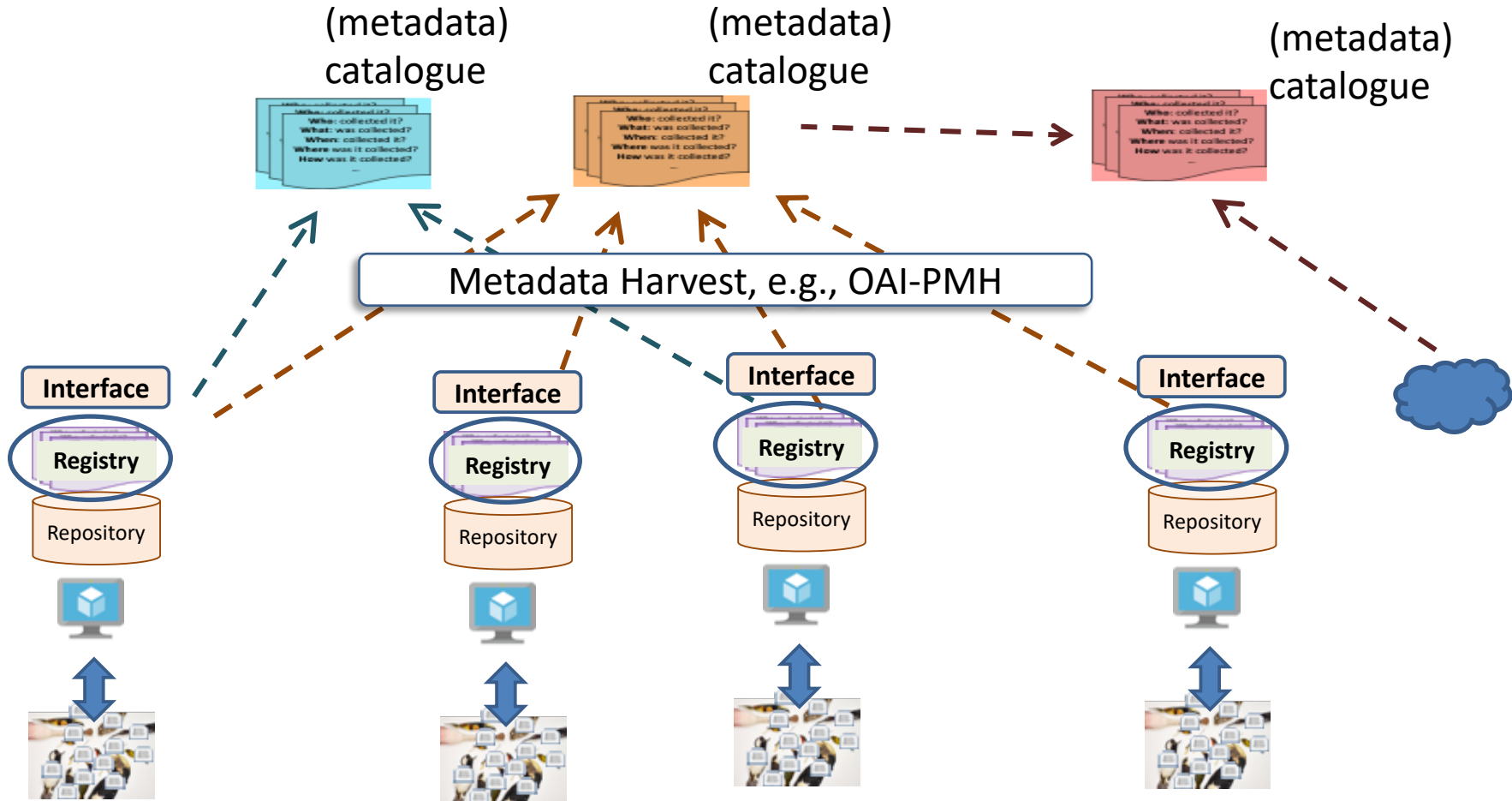


} For administrator, for user, and for other services (catalogues, workflows, search engines, etc.)

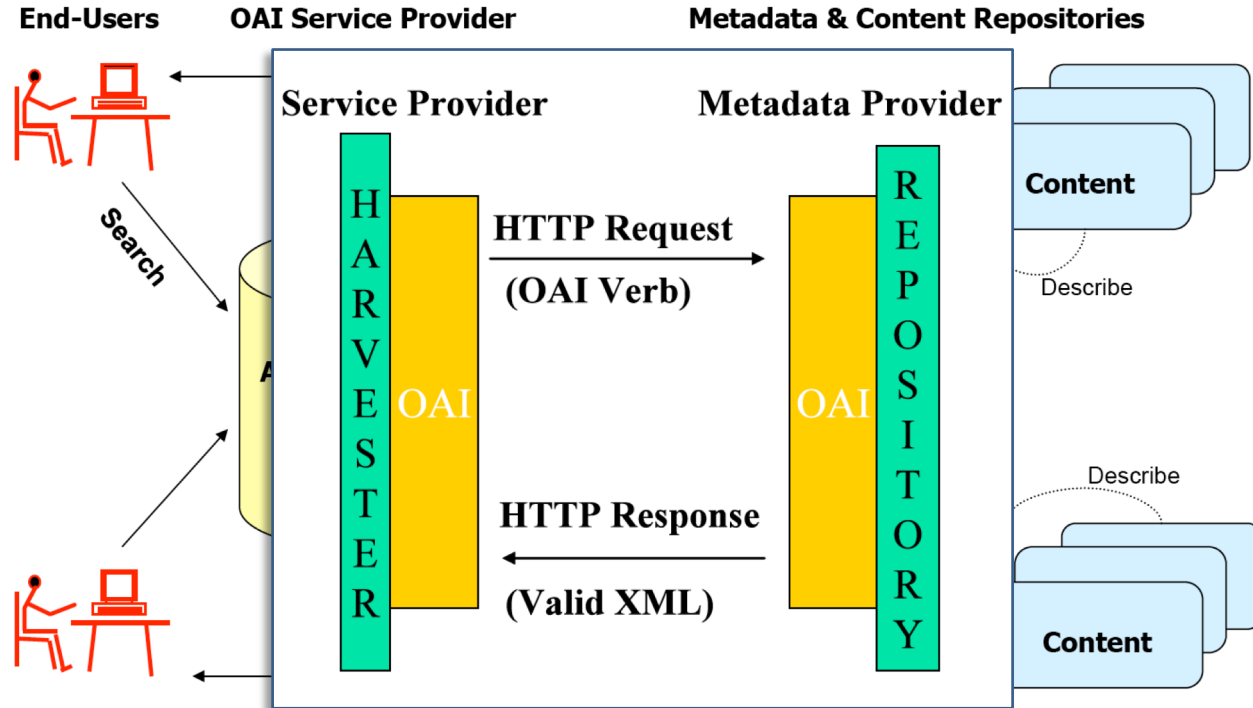
Discussion

- How are catalogue and PID related?

Catalogue for publishing and discovery

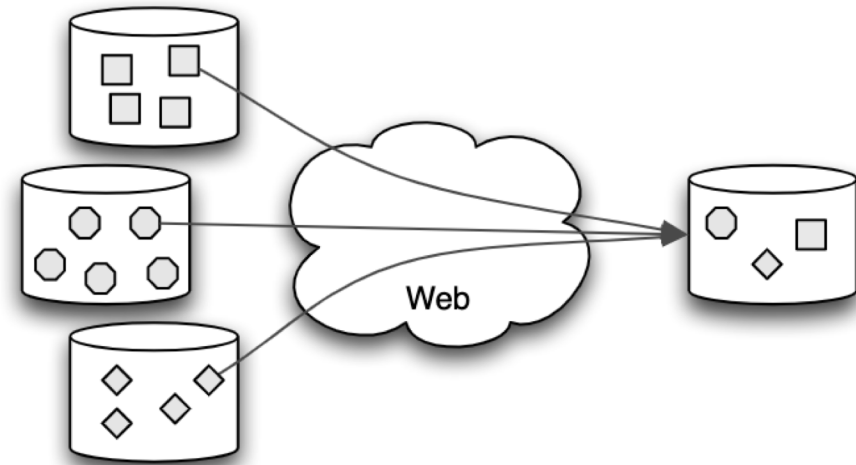


How does harvest work?

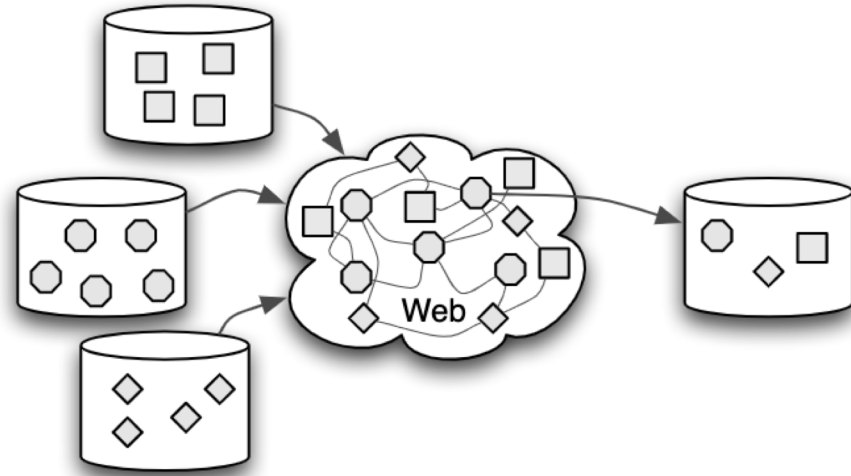


Other solutions

- Linked Open data approach



OAI-PMH Approach



LOD Approach

Discussion

- Does catalogue mean open access?

Discussion

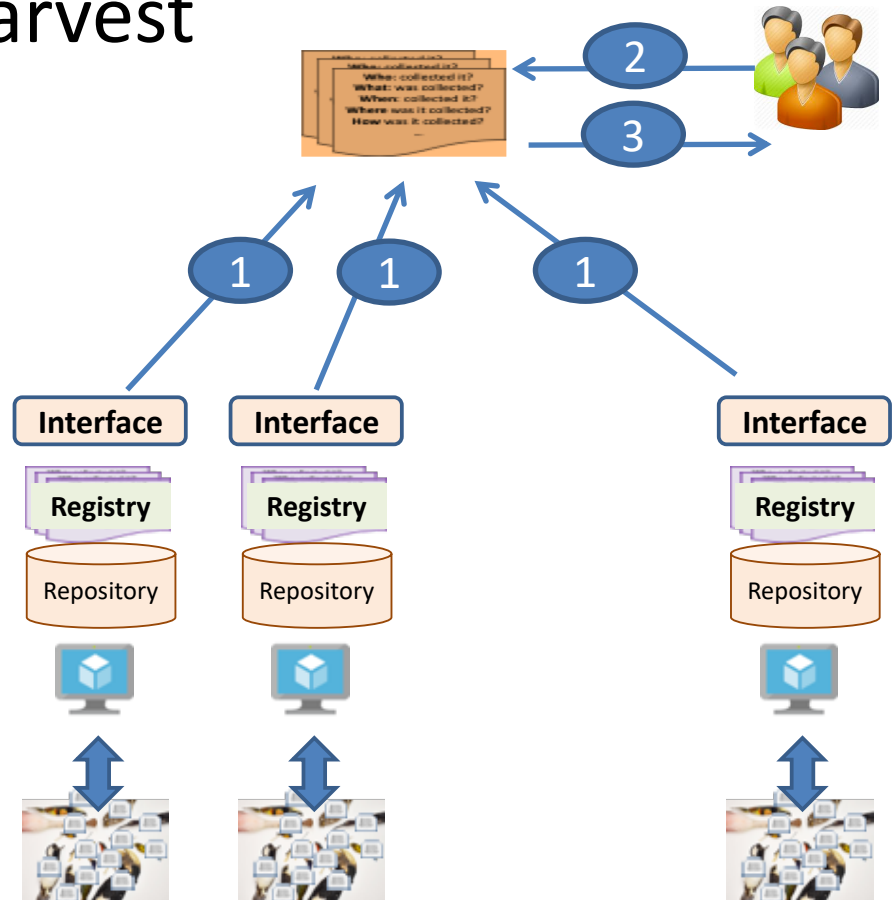
- Does catalogue mean open access?
- Catalogue is effective for:
 - Findability
 - Accessibility
 - Interoperability
 - Reusability

Discussion: how to search across catalogue

- Approaches:

Approach one: catalog harvest

- Harvest distributed catalogues
- Query the harvested catalogue



Cross RI

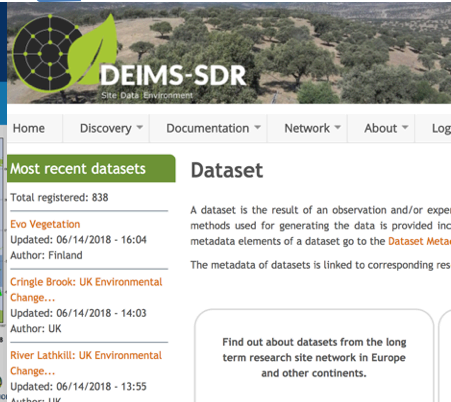
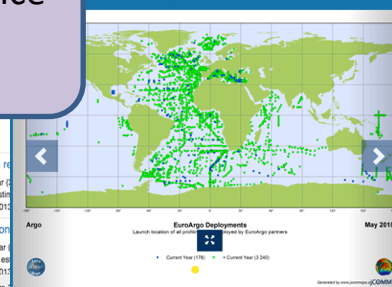
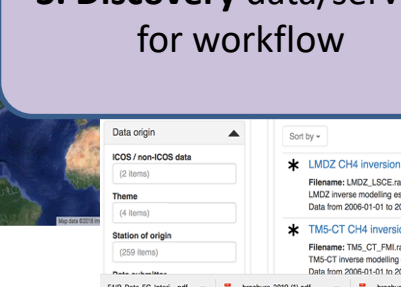
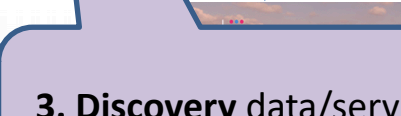
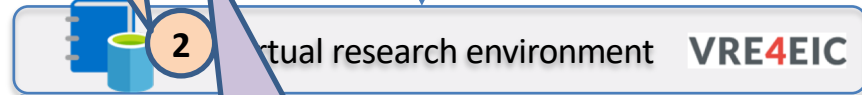
RI using

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

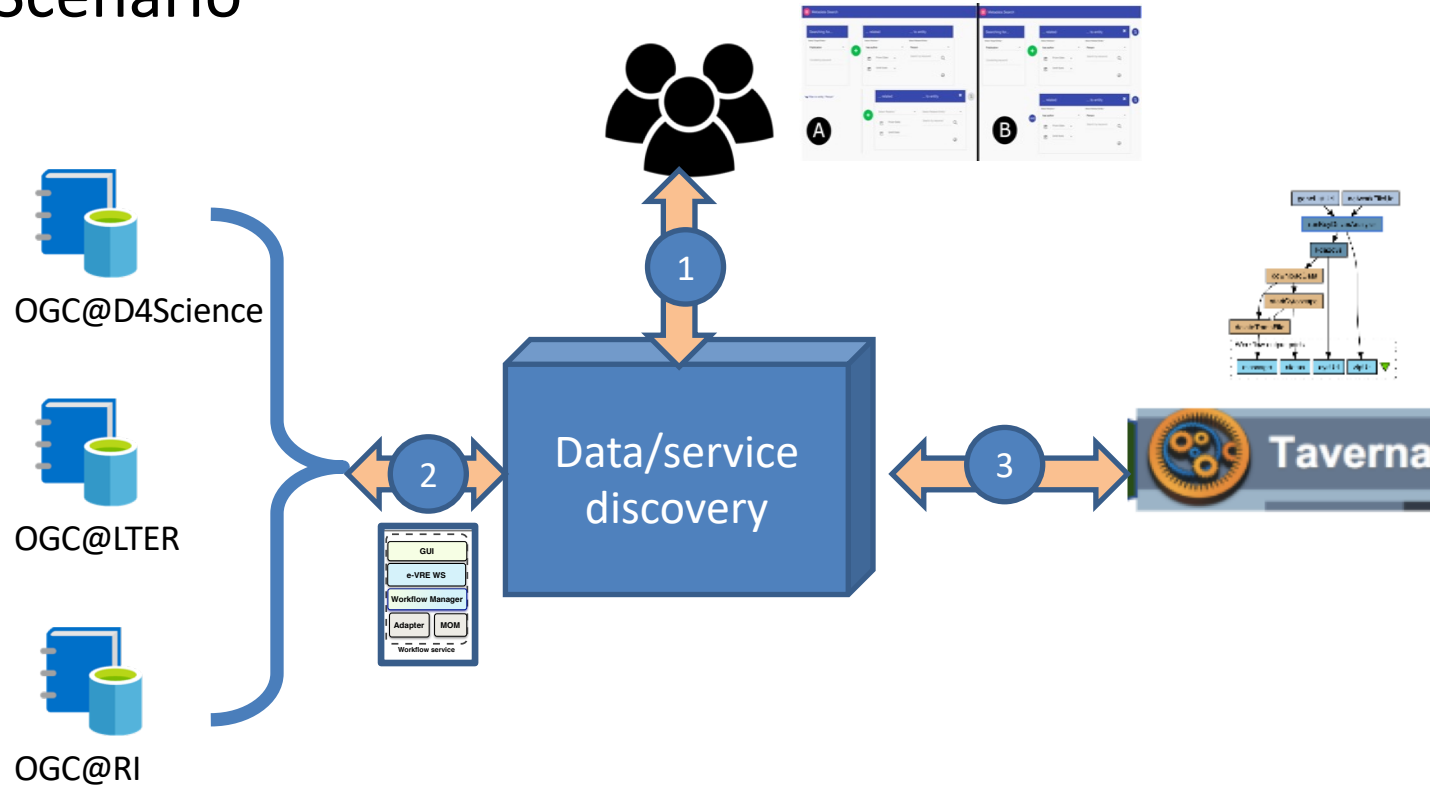
4. Execute workflow and manage scientific experiments (Workflow manager/etc.)

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

3. Discovery data/service for workflow

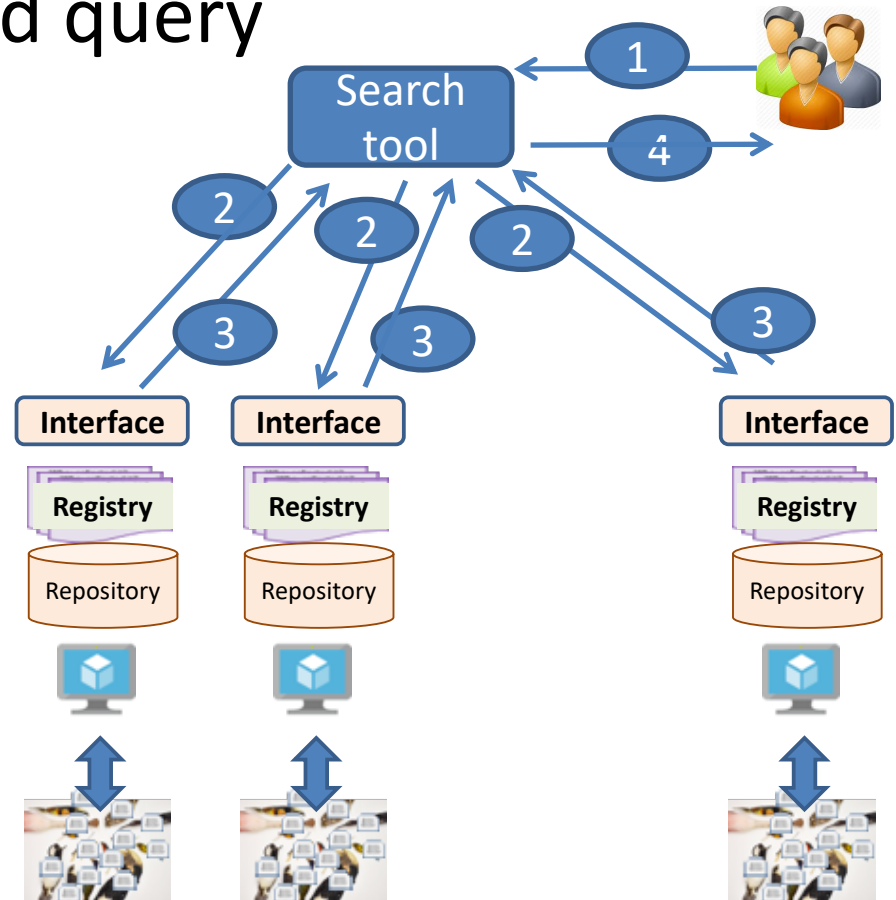


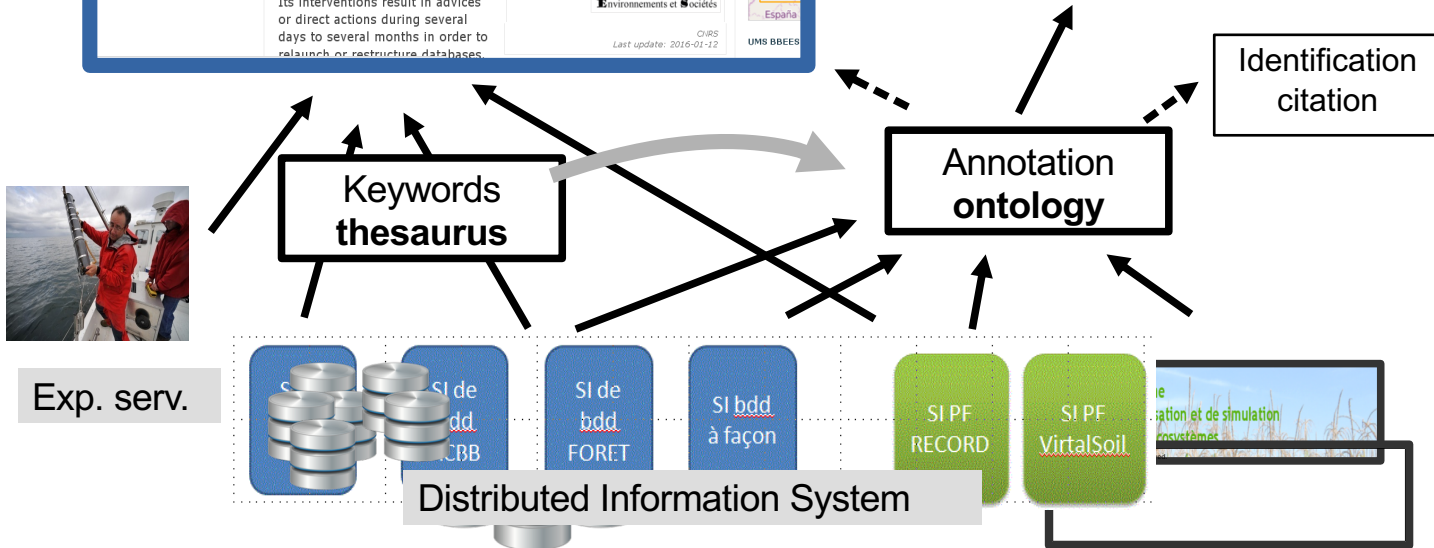
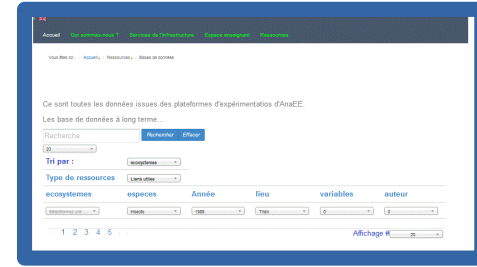
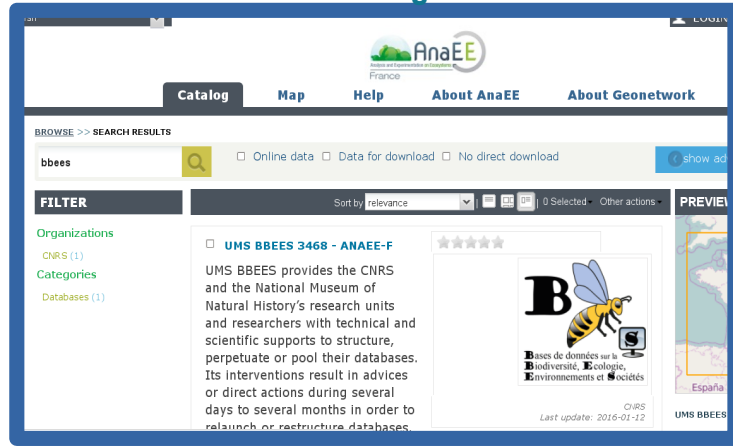
Scenario



Approach two: distributed query

- Generate queries to distributed catalogues
- Collect the results





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