Proiectul: *Organizarea bazei de date a INCD GeoEcoMar pentru operaționalizarea conceptului de Fair Data*

Etapa1: *Analiza standardelor si a schemelor precum si a tipurilor de date existente in cadrul INCD GEOECOMAR (termen predare:* **14.06.2023)**

Activitati derulate Etapa 1:

- ***Analiza tipurilor de date existente in cadrul bazei de date actuale a INCD GEOECOMAR***

In prezent datele existente in baza de date a institutului sunt catalogate in funcție de specificul lor pe 5 direcții principale de cercetare: geologie, geofizica, biologie, hidrologie si oceanografie. Acestea la rândul lor sunt împărțite pe subdirecții de cercetare specifice fiecărui domeniu principal conform figurii 1:

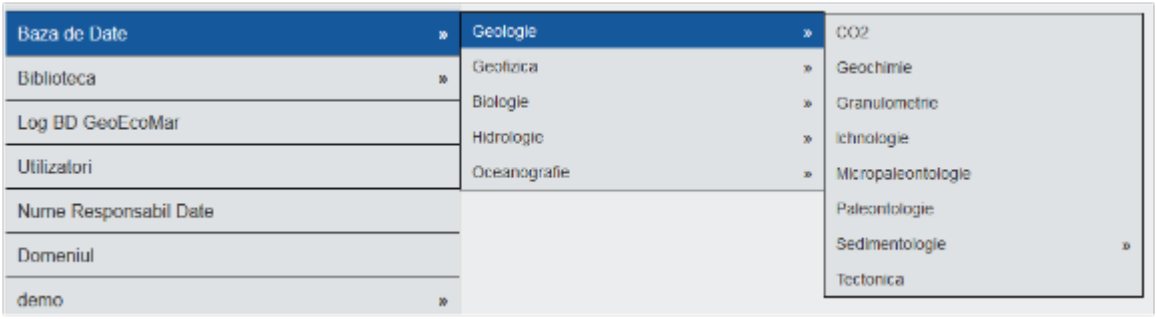

Figura 1 – structura bazei de date pe baza meniului de acces

Tipurile de fișiere încărcate au următoarele tipuri de extensii:

- fișiere text – „.txt”
- fișiere Excel – „.csv”, „.doc”, „.docx”, „.xls”, „.xlsx”
- arhive rar si zip cu fișiere specifice exportate in urma măsurătorilor de teren

Datele sunt încărcate folosind un formular cu cerințe definite clar pentru a permite căutarea si refolosirea datelor.

- ***Analiza celor mai populare standarde si scheme existente la nivelul actual***

În documentarea datelor cu metadate, este important să respectam standardele - vocabulare standard, scheme standard etc. Organizarea datelor folosind o abordare bazată pe standarde ajută la asigurarea interoperabilității între sisteme, ceea ce îmbunătățește, de asemenea, descoperirea și accesul la date.

A fost făcută o analiza a celor mai relevante standarde si scheme folosite urmărindu-se adaptarea soluțiilor observate in cadrul proiectului pentru obținerea unui „Fairness” cit mai adecvat

Rezultate Etapa1:

- ***Elaborarea unui set de principii specifice de adaptare a datelor si metadatelor pentru obținerea unui portal de date FAIR***

In cadrul acestei etape a fost analizata baza de date actuala a INCD GEOECOMAR. Pe baza observațiilor apărute in urma analizei datelor introduse, a metadatelor definite si respectarea principiilor FAIR de către baza de date actuala tragem concluzia ca actuala baza de date respecta doar parțial unele dintre principiile FAIR Data Data:

Criteriu FAIR Data	**Grad de indeplinire**
„găsibil”	acest principiu este respectat parțial - in special trasabilitatea datelor cerându-se completarea următoarelor metadate definite: nume responsabil date, expediția, instrument achiziție date, data (achiziției), locație
„accesibil”	acest principiu este respectat parțial datele fiind accesibile din exterior dar numai către personalul institutului pe baza unei cerere de acces. Nu exista posibilitatea interogării datelor folosind limbaje de programare si scripturi dedicate
„refolosibil”	acest principiu este respectat parțial. Exista o descriere a instrumentelor folosite pentru achiziția datelor, formatul de date si se oferă opțional posibilitatea unor comentarii asupra datelor. Nu exista definite licențe de acces
„interoperabil”	acest principiu nu este respectat de baza de date existenta.

De asemenea s-a făcut o analiza pe principalele standarde si scheme folosite la ora actuala si pe baza acestora elaboram următoarele principii pentru obținerea unui portal de date cu un „Fairness” cit mai ridicat:

Criteriu FAIR Data	**Set de principii elaborat pentru indeplinire criteriu**
„găsibil”	- Generarea de DOI (digital object identifier) pentru seturile de date - Crearea unor API-uri de acces la seturile de date. - Descrierea datelor folosind un catalog cit mai complet de metadate folosind definiții din schemele existente si analizate in aceasta faza: DataCite, Dublin Core, DCAT
„accesibil”	- Folosirea de DOI pentru seturile de date - Crearea unor licențe de acces la seturile de date si a unui mecanism de înregistrare si drepturi de utilizator cit mai bine reglementat - Acces liber la metadate chiar daca datele sunt inaccesibile
„refolosibil”	Documentare completă a seturilor de date: instrumente folosite, autor date, proceduri, licențe acces.

„interoperabil”	- Stocarea de seturi de date care sa poată fi vizualizate si prelucrate cu softuri open source acolo unde este posibil - Folosirea unor standarde relevante pentru metadatele descriptive ale seturilor de date care vor fi stocate in portal

Diseminare Etapa 1:

- *METHODS USED TO ADAPT FAIR PRINCIPLES TO SCIENTIFIC DATA Cristian CUDALBU, Teodora BABOS, Florin PITEA – GeoEcoMarina 29/2023*

Echipamente Etapa 1:

- PC analiza/modelare date CreativeX NVIDIA Studio Render (Intel i9-13900K 3.0GHz, 64GB DDR5, 2TB SSD, RTX 4090 24GB GDDR6X)

Etapa2: *Analiza celor mai populare portaluri de date existente (termen predare:* **07.12.2023)**

Activitati derulate Etapa 2:

- ***Instalarea, configurarea si analiza candidaților pentru portalul de date***

  In urma analizei specificului de date existente in cadrul INCD GEOECOMAR desfășurata in faza 1 a proiectului s-a concluzionat ca acestea sunt variate, pe diverse direcții de cercetare fiecare cu specificul ei in afișarea criteriilor relevante si tipul de fișiere obținute. Acest domeniu variat al datelor face ca alegerea unui portal de date sa fie o sarcina complexa fiind necesara instalarea de proba a acestora si analiza in funcție de mai multe criterii pentru a desemna cu succes cea mai potrivita soluție pentru specificul datelor existente in cadrul GeoEcoMar.

  La acest moment exista numeroase soluții open source de portaluri de date atât self hosted (instalate pe un server in cadrul instituției) cit si cloud hosted (instalate pe serverele altei organizații).

  Printre cele mai folosite tipuri de portaluri de date existente la ora actuala se numără: dataverse, CKAN, Geonetwork, DKAN.

  CKAN a fost ales în principal pentru că este unul dintre cele mai mari

  platforme de date deschise de pe piață, este bine documentată, inclusiv instrucțiuni de instalare și acceptă orice format de fișier ca sursă de date. Geonetwork a fost ales pentru ca este unul dintre cele mai folosite portaluri de date pentru date cu referința geospațiala dar nu numai. Dataverse si DKAN au fost alese pentru ca sunt folosite de un număr ridicat de instituții publice si au caracteristici funcționale adecvate pentru obiectivul proiectului.

  Dataverse, CKAN, DKAN si Geonetwork au fost instalate pe mașini virtuale si testate cu date sample.

  Au fost făcute observații privind următoarele criterii:

- ușurința in instalare si configurare
- arhitectura
- caracteristici de baza, extensii

Rezultate Etapa2:

- ***instalare si testare cele mai populare portaluri de date existente pe o mașină virtuală.***
- ***desemnare candidat portal de date al proiectului***

In cadrul acestei etape au fost instalate si configurate cele mai populare portaluri de date open source folosite de către instituții guvernamentale si entități de cercetare științifica. De asemenea au fost încărcate seturi de date test si urmărite funcționalitățile de baza ale acestor portaluri precum si posibilitatea extinderii acestora folosind extensii sau plugine oferite de către dezvoltatorii acestor portaluri.

Portalurile testate au fost caracterizate din punct de vedere al ușurinței in instalare si configurare, arhitectura, caracteristici de baza - extensii .

In urma analizei efectuate s-a creat un tabel in care cei patru candidați au primit punctaj de la 1 la 5 pentru:

- formate sau tipuri de date care pot fi încărcate
- tipuri fluxuri ieșire date (descărcare, preview, API)
- ușurința in operare de către utilizatori
- calitate documentație
- instrucțiuni instalare si configurare
- compatibilitate FAIR Data

Rezultate:

- CKAN: 29 puncte
- Dataverse: 25 puncte
- Geonetwork: 25 puncte
- DKAN: 23 puncte

Analizind rezultatele de mai sus CKAN a obținut cel mai mare punctaj si a fost desemnat ca soluția de portal cea mai potrivita pentru nevoile bazei de date INCD GEOECOMAR aceasta urmând a fi instalata si configurata in etapele viitoare ale proiectului conform schemei de realizare.

Diseminare Etapa 2:

-

Echipamente Etapa 2:

- Laptop analiza/modelare date Lenovo Pro 7 16IRX8H, WQXGA IPS 240Hz G-Sync, Procesor Intel® Core™ i9-13900HX (36M Cache, up to 5.40 Ghz, 32GB DDR5, 1TB SSD, GeForce RTX 4090)