FAIRificarea datelor de cercetare - etape și evaluare

Evenimentul

”National End-user Training”

29.03.2022

Alexandru Stanciu, Gabriel Neagu
ICI București

Inițiative Naționale pentru Știință Deschisă în Europa – H2020 Research and Innovation action – contract no. 857645

Planul prezentării

- Context
  - Date de cercetare deschise
  - Principiile FAIR
- Metrici FAIR
- FAIRificarea datelor - un proces evolutiv
  - O propunere de structuare a procesului
- Relația DMP – procesul de FAIRificare
- Soluții de evaluare a nivelului FAIR al datelor

Date de cercetare deschise

- *Directiva UE privind datele deschise*, iunie 2019:
  - **Date de cercetare**: *"documente în format digital care nu sunt publicații științifice și care sunt colectate sau produse în cursul activităților de cercetare științifică și sunt utilizate ca dovezi în procesul de cercetare sau sunt acceptate în mod curent în comunitatea de cercetare drept necesare pentru validarea concluziilor și rezultatelor cercetărilor"*.
  - **Accesul deschis**: *"practica de a furniza acces online la rezultatele cercetării cu titlu gratuit pentru utilizatorii finali, cât mai devreme posibil în cadrul procesului de difuzare și fără restricții de utilizare și reutilizare, dincolo de posibilitatea de a solicita recunoașterea calității de autor"*.
  - **Beneficii ale datelor de cercetare deschise**:
    - îmbunătățirea calității
    - reducerea duplicării inutile a cercetărilor
    - accelerarea progresului științific
    - combaterea fraudelor în domeniul științei
    - favorizarea inovării și a creșterii economice în general

Principiile FAIR

- Caracterizare
  - sunt principii directoare care descriu calitățile necesare **pentru ca datele să fie maxim reutilizabile**
  - ghidează dezvoltarea infrastructurii și instrumentelor care susțin obiectele de cercetare să devină reutilizabile în mod optim, **pentru mașini și oameni deopotrivă**
  - se pot aplica datelor indiferent de disponibilitatea publică a acestora și, în mod specific, **nu necesită ca datele să fie deschise**
  - având în vedere obiectivul reutilizării, prioritară este aplicarea lor pentru **datele create de cercetarea finanțată public**
  - reprezintă **un set de linii directoare esențiale (de aur) pentru managementul datelor de cercetare (RDM)**; asigurarea caracterului FAIR la datelor este o precondiție esențială pentru un RDM adecvat

Date Open & FAIR

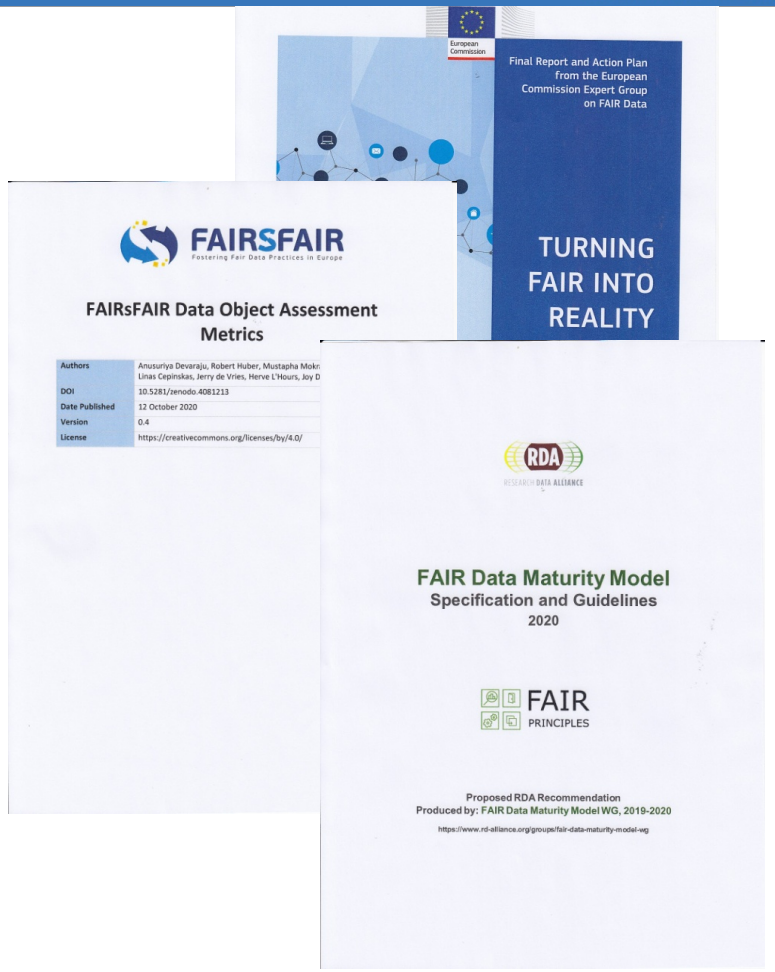
- *RECOMANDAREA (UE) 2018/790 A COMISIEI privind accesul la informațiile științifice și conservarea acestora*:
  - *”statele membre ar trebui să stabilească și să pună în aplicare **politici clare pentru managementul datelor cercetării rezultate din cercetarea finanțată din fonduri publice**, prin care să se asigure inclusiv de faptul că **datele din această categorie devin și rămân ușor de găsit, accesibile, interoperabile și reutilizabile („principiile FAIR”)** într-un mediu sigur și fiabil, prin infrastructuri digitale [**inclusiv cele grupate în Cloudul european pentru știința deschisă (EOSC), după caz**], cu excepția cazului în care acest lucru nu este posibil sau nu este compatibil cu exploatarea în continuare a rezultatelor cercetării (**cât mai deschis posibil, atât de închis cât este necesar**)”.*
- *Regulamentul (UE) 2021/695 de instituire a programului-cadru Orizont Europa*:
  - **accesul deschis la datele de cercetare**, inclusiv la cele care stau la baza publicațiilor științifice, în conformitate cu principiul „cât mai deschis posibil, atât de închis cât este necesar” și **managementul responsabil al datelor în conformitate cu principiile FAIR** reprezintă modalități de sprijin al științei deschise.

Cerințe FAIR

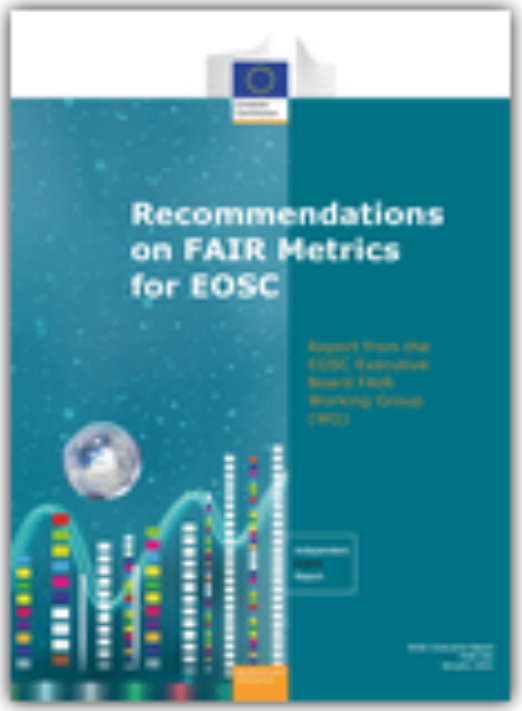
- ***Findable***:
  - F1. (meta)datele au asignat un identificator unic la nivel global și persistent
  - F2. datele sunt descrise prin intermediul metadatelor
  - F3. (meta)datele sunt înregistrate sau indexate pe suport care permite căutarea
  - F4. Metadatele specifică clar și explicit identificatorul datelor pe care le descriu
- ***Accesible***:
  - A1. (meta)datele pot fi regăsite pe baza identificatorului lor folosind un protocol de comunicații standardizat:
    - A1.1. protocolul este deschis, gratuit și universal implementabil (ex. HTTP, FTP)
    - A1.2. este definită, după caz, o procedură de acces
  - A2. (meta)datele sunt accesibile chiar și după ce datele nu mai sunt disponibile
- ***Interoperable***:
  - I1. (meta)datele utilizează un limbaj formal, accesibil, partajat și aplicabil pe scară largă pentru reprezentarea cunoștințelor
  - I2. (meta)datele folosesc vocabulare, tezaure, ontologii care respectă principiile FAIR
  - I3. (meta)datele includ referințe calificate la alte (meta)date
- ***Reusable***:
  - R1. (meta)datele au o multitudine de atribute precise și relevante
  - R1.1. (meta)datele sunt eliberate cu o licență clară și accesibilă de utilizare
  - R1.2. (meta)datele sunt asociate cu proveniența lor
  - R1.3. (meta)datele respectă standardele comunității relevante din domeniu

Metrici FAIR

- **Turning FAIR into Reality** – Final report and action plan from the EC Expert Group on FAIR Data (2018)
  - **componentele "realității FAIR"**: obiecte FAIR, ecosistem FAIR (servicii, specificații metadate, depozite digitale, DPM), suport de interoperabilitate, FAIR accesibil pentru om și calculator, competențe Data science și Data stewardship, metrici și indicatori de implementare, corelați cu stimulente
- RDA **FAIR Data Maturity Model** – Specification and Guidelines (iunie 2020)
  - 41 de indicatori: 7 - Findable, 12 - Accesible, 12 -Interoperable, 10 - Reusable
- FAIRsFAIR **Data Object Assessment Metrics** (oct. 2020)
  - 17 metrici: 5 – Findable, 4 – Accesible, 3 – Interoperable, 5 – Reusable



- **Recommendations on FAIR Metrics for EOSC**, EOSC Executive Board FAIR Working Group, 2021
  - 6 recomandări privind definirea și implementarea metricilor
  - un set de 25 metrici-țintă pentru datele FAIR în EOSC, *derivate din RDA FAIR Data Maturity Model*, implemetate conform unei planificări în 3 etape (2021 / 2024 / 2028)
  - 7 priorități în definirea metricilor FAIR, considerat un proces continuu

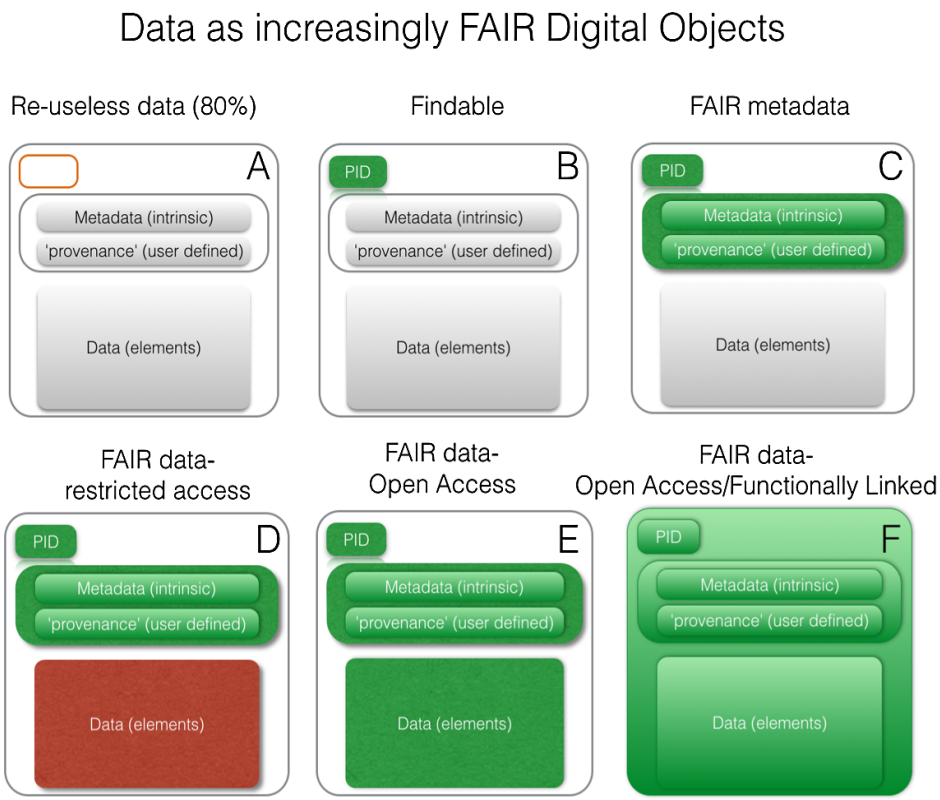


FAIRificarea datelor - un proces evolutiv

- FAIRsharing:
  - *”vedem FAIR ca **un continuum de „comportamente”** manifestate de o resursă de date care permite progresiv descoperirea și (re)utilizarea de către calculator”*
  - „FAIR” va avea **cerințe diferite pentru comunități diferite**
- RDA – FAIR Data Maturity Model:
  - **5 niveluri de FAIRificare**, funcție de seturile de indicatori FAIR implementați
    - Nivel 1-indicatorii esențiali (19)
    - Nivel 2-indicatorii esențiali + 50% indicatori importanți (26)
    - Nivel 3-indicatorii esențiali + 100% indicatori importanți (34)
    - Nivel 4-indicatorii esențiali + 100% indicatori importanți + 50% indicatori utili (37)
    - Nivel 5-indicatorii esențiali + 100% indicatori importanți + 100% indicatori utili (41)
- Recommendations on FAIR Metrics for EOSC:
  - **FAIRificarea ar trebui considerată o călătorie**: implementare treptată, cu instrumente de evaluare pentru măsurarea progresului

Partly FAIR may be fair enough [1]

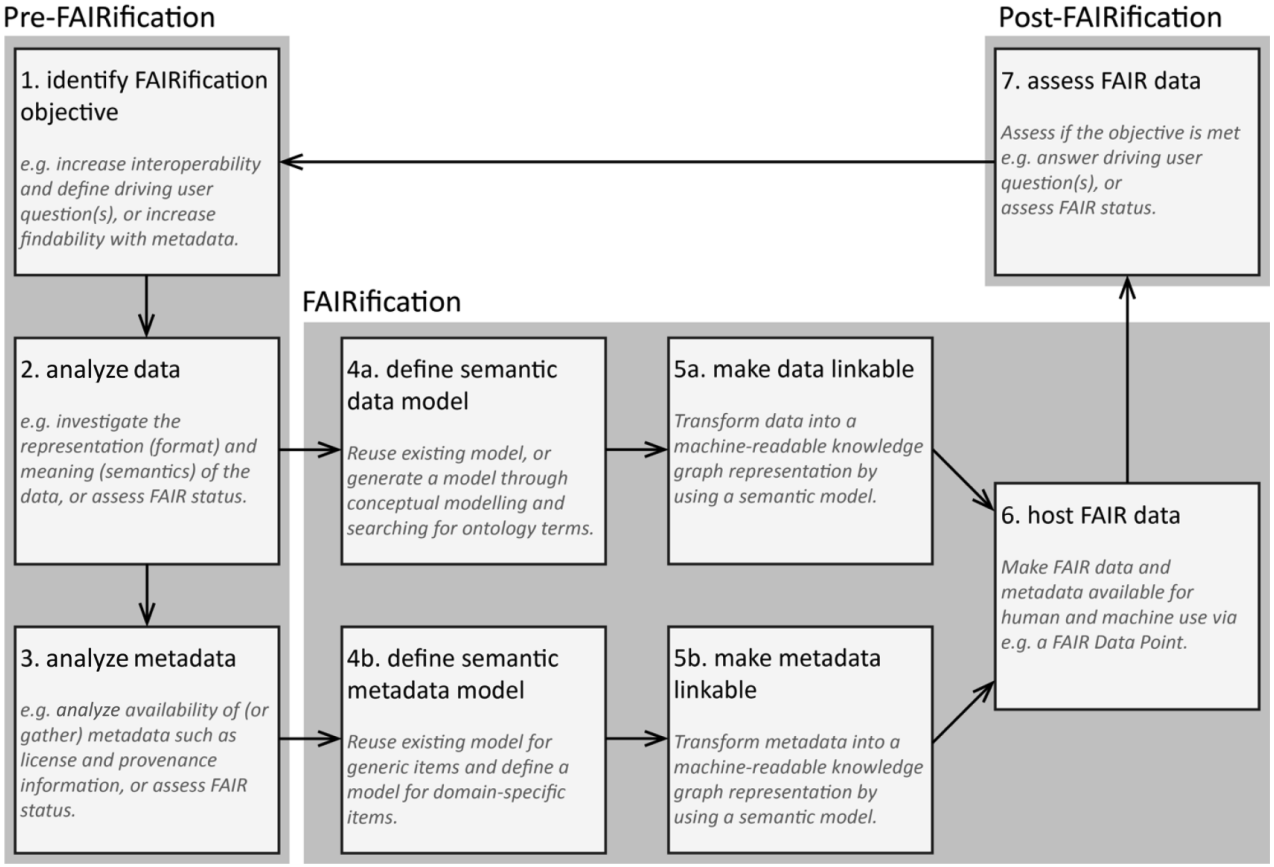
- A: date indisponibile pentru reutilizare
- B: nivel minim de FAIRifcare - furnizarea setului de date, ca entitate în sine, cu un PID care nu este doar intrinsec persistent, ci și legat persistent de setul de date
- C: atât metadatele „intrinseci” cât și metadatele „definite de utilizator” ar trebui să fie adăugate și FAIRificate ori de câte ori este posibil
- D: datele sunt FAIR din punct de vedere tehnic, dar din diverse motive este necesară restricționarea accesului
- E: nivelul maxim de FAIRficare - datele în sine sunt disponibile pentru reutilizare de către alții, în condiții bine definite
- F: Internetul datelor și serviciilor FAIR - un număr în creștere de aplicații și servicii pot găsi și procesa date FAIR



O propunere de structurare a procesului [2]

- Trei etape:
  - pre-FAIRificare / FAIRificare / post-FAIRificare
- Șapte pași:
  - Identificarea obiectivelor pentru FAIRificare
  - Analiza datelor
  - Analiza metadatelor
  - Definirea unui model semantic pentru date și metadate
  - Asigurarea asociativității datelor și metadatelor
  - Publicarea datelor FAIR
  - Evaluarea nivelului de FAIRificare a datelor
- Nu este obligatoriu să se respecte o ordine strictă
  - Seturile de date pot avea anumite caracteristici sau constrângeri de ordin practic care determină o altă ordine de execuție a procesului de FAIRificare sau repetarea unor pași.
  - Fiecare pas are drept scop eficientizarea implementării principiilor FAIR și îmbunătățirea scorului FAIR (FAIRness) al setului de date.

Procesul de FAIRificare a datelor



Un flux de lucru generic, pas cu pas, pentru procesul de FAIRificare (Sursa: [2])

Competențe necesare pentru FAIRificare

- Echipe multi-disciplinare
- Exemple de domenii de expertiză:
  - Domeniul propriu zis al datelor care urmează a fi FAIRificate
  - Modul de gestionare al datelor
  - Arhitectura instrumentelor specifice pentru gestionarea datelor
  - Definirea de politici de acces aplicabile la datele în cauză
  - Coordonarea și monitorizarea procesului de FAIRfificare
  - Implementarea instrumentelor sau a serviciilor necesare pentru FAIRificare
  - Modelarea datelor
  - Standarde specifice resurselor de tip date și accesării datelor

Etapa pre-FAIRificare

Pasul 1 - Definirea obiectivelor

- Este necesar să existe o imagine generală a setului de date.
- Este necesar accesul la date
  - În situația în care datele au un caracter privat ”sensitive” și nu este posibil accesul nelimitat la date se poate utiliza un subset anonimizat sau alte date fictive.
- Se recomandă să se ajusteze volumul de date utilizat cu resursele disponibile pentru FAIRificare (ex. timp).
  - Procesul poate fi repetat de mai multe ori, astfel încât datele suplimentare pot fi adăugate ulterior.
- În general, obiectivele pentru FAIRificare depind de: a) expertiza disponibilă, b) soluțiile sau instrumentele FAIR care pot fi reutilizate, c) instrumente care suporta FAIRficarea datelor și care sunt proiectate pentru tipul de date în cauză.

Etapa pre-FAIRificare

Pasul 2 – Analiza datelor

- Pregătrea următorilor pași din procesul de FAIRificare (ex. îmbunătățirea interoperabilității datelor)
- Activități posibile:
  - Investigarea modurilor în care pot fi reprezentate datele și verificarea dacă modul de reprezentare (formatul) și semantica datelor sunt clare și neambigui.
  - De investigat dacă nu există deja unele caractisici FAIR (ex. identificatori unici pentru date) prin utilizarea unor instrumente pentru evaluare (data FAIRness).

Etapa pre-FAIRificare

Pasul 3 – Analiza metadatelor

- Analiza disponibilității metadatelor pentru regăsirea, accesibilitatea și reutilizarea datelor
  - Datele sunt făcute a fi interoperabile în etapele ulterioare (definirea unui model semantic și asocierea cu metadate)
- Activități:
  - Identificarea unor metadate care descriu cât mai bine datele respective
  - Verificarea dacă aceste metadate nu conțin deja elemente FAIR (ex. descrierea sursei datelor) utilizând instrumente pentru evaluarea (data FAIRness).
  - Îmbunătățirea metadatelor referior la caracteristicile de regăsire, accesibilitate, reutilizabilitate necesită includerea unor informații precum: **licență, copyright, contribuție (finanțare, creare, publicare), condiții de acces și utilizare**.

Etapa FAIRificare

Pasul 4 – Definirea unui model semantic (date și metadate)

- Modelele semantice reprezintă șabloane pentru transformarea datelor și metadatelor (asocierea acestora pentru a asigura interoperabilitatea datelor) într-un format specific procesării automate.
- Este cel mai complex pas și necesită cel mai mare efort în procesul de FAIRificare
  - Din ce în ce mai multe modele semantice sunt dezvoltate și pot fi reutilizate -> de verificat dacă nu există un astfel de model semantic pentru date și metadate.
- Procesul de implementare al unui model semantic include 3 activități succesive:
  1) Crearea unui model conceptual
  2) Identificarea unei ontologii pentru termenii utilizați
  3) Construirea modelului semantic pe baza elementelor (1) și (2).
- Sunt necesare cunoștințe referitoare la modelarea semantică a datelor

Definirea unui model semantic

- Modelul semantic al datelor se realizează în baza modelului conceptual și a ontologiei utilizate
  - Spre deosebire de modelul conceptual, modelul semantic specifică atât datele (instanțele și valorile acestora), cât și tipul acestora (clasa din care fac parte).
  - Este o reprezentare exactă a datelor și expune sensul acestora într-un format care poate fi procesat cu ușurință de un program (ideal ar fi un format universal)
  - Permite transformarea datelor FAIR astfel încât să poată fi incluse în alte sisteme, aplicații sau fluxuri de lucru (procese)
  - În cazul metadatelor, modelarea semanatică prezintă elementele generice care ar putea fi reutilizate (ex. DCAT pentru a descrie un set de date într-un catalog).

Etapa FAIRificare

Pasul 5 – Asigurarea asociativității datelor și metadatelor

- Reprezintă un proces specific fiecărui domeniu sau aplicații.
- Este esențial să existe o reprezentare a datelor și metadatelor într-un format specific mașină (ex. Resource Description Framework - RDF).
- Se utilizează modelul semantic al datelor
  - Instrumente specializate: FAIRifier, Karma, Rightfield, OntoMaton.
- Pentru transformarea metadatelor într-un format specific mașină se utilizează modelul semantic al metadatelor
  - În cazul unor elemente generice se pot utiliza instrumente precum: FAIR Metadata Editor, CEDAR, BioSchemasGenerator.

Etapa FAIRificare

Pasul 6 - Publicarea datelor FAIR

- Asigură disponibilitatea și posibilitatea de utilizare a acestora de către cei interesți (conform cu licența utilizare, politici de acces etc.)
  - Atât persoane cât și aplicații
  - Definirea și utilizarea unor interfețe specifice (API), RDF triplestore, servicii web etc.

Etapa post-FAIRificare

Pasul 7 - Evaluarea datelor FAIR

- Activități propuse:
  - Verificarea atingerii obiectivelor definite la pasul 1 (în cazul în care nu fost atinse toate obiectivele se pot relua anumiți pași).
  - Verificarea nivelului FAIR (FAIRness) al datelor și metadatelor cu ajutorul unor instrumente specifice

Relația DMP – procesul de FAIRificare [3]

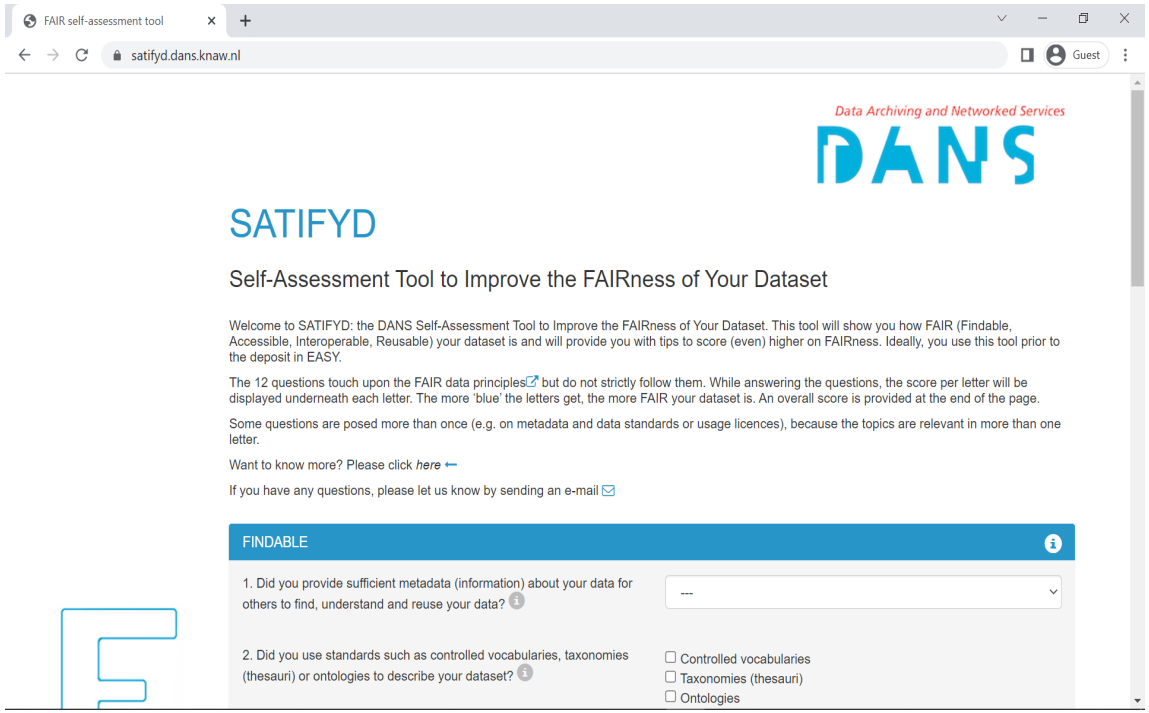
- Raportul *Practical Guide to the International Alignment of Research Data Management* elaborat Science Europe:
  - Șase cerințe de bază pentru DMP, detaliate în 15 interogări
  - Ghid privind elaborarea DPM pe baza acestor cerințe
  - Ghid de evaluare a DMP raportat la aceste cerințe
  - Corespondența între cerințele FAIR (v. slide 6) și cele 15 interogări
- Rezultă o posibilitate de structurare a procesului de FAIRificare pe secțiunile DMP
  - Ordinea în care diversele cerințe FAIR sunt abordate în diverse etape ale DMP.

O structurare a FAIRificării pe suport DMP

Six core requirements for good data management	FAIR Requirements
1 DATA DESCRIPTION AND COLLECTION OR RE-USE OF EXISTING DATA	
1a - How will new data be collected or produced and/or how will existing data be re-used?	R1.2
1b - What data (for example the kind, formats, and volumes) will be collected or produced?	I1, R1.2, R1.3
2 DOCUMENTATION AND DATA QUALITY	
2a - What metadata and documentation (e.g. the methodology of data collection and way of organising data) will accompany the data?	F2, I1, I2, I3, R1
2b - What data quality control measures will be used?	I2, R1
3 STORAGE AND BACKUP DURING THE RESEARCH PROCESS	
3a - How will data and metadata be stored and backed up during the research?	
3b - How will data security and protection of sensitive data be taken care of during the research?	
4 LEGAL AND ETHICAL REQUIREMENTS, CODES OF CONDUCT	
4a - If personal data are processed, how will compliance with legislation on personal data and on security be ensured?	
4b - How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?	A1.2, R1.1
4c - What ethical issues and codes of conduct are there, and how will they be taken into account?	A2
5 DATA SHARING AND LONG-TERM PRESERVATION	
5a - How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?	A1.2, A2, I3, R1.1
5b - How will data for preservation be selected, and where data will be preserved long-term (e.g. a data repository or archive)?	
5c - What methods or software tools are needed to access and use data?	A1.1, A1.2, I3
5d - How will the application of a unique and persistent identifier (such as DOI) to each data set be ensured?	F1, F3, A2
6 DATA MANAGEMENT RESPONSIBILITIES AND RESOURCES	
6a - Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)?	
6b - What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR?	

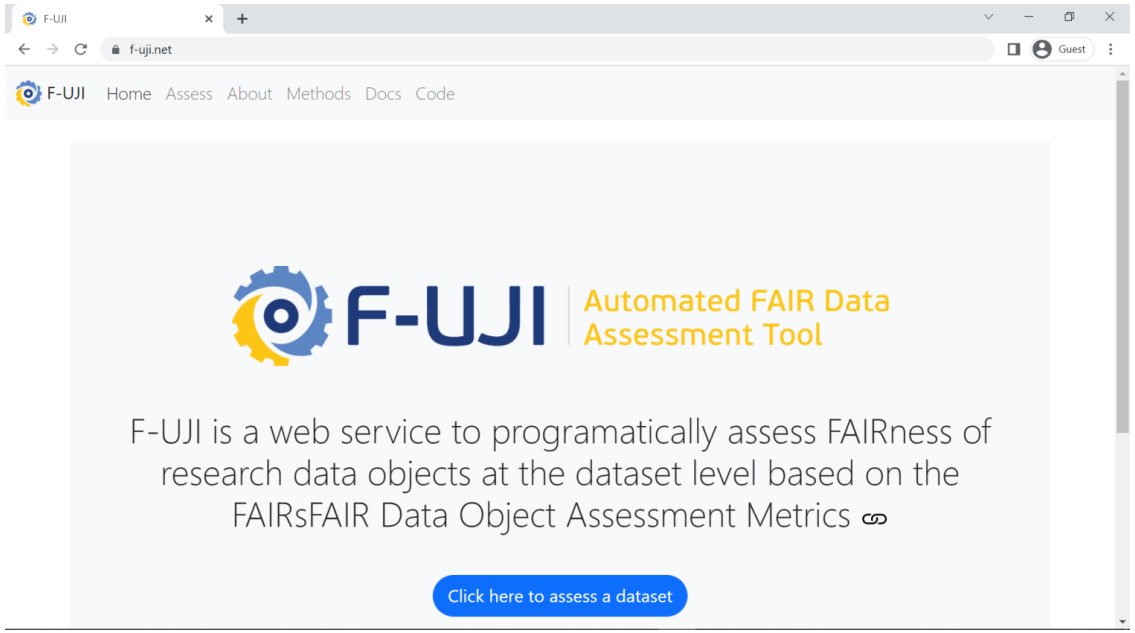
Soluții de evaluare

- SATIFYD - Self-Assessment Tool to Improve the FAIRness of Your Dataset (DANS)
  - Chestionar online de 12 intrebări care acopera cele 4 principii FAIR
  - Furnizează în final un scor al nivelului FAIRness



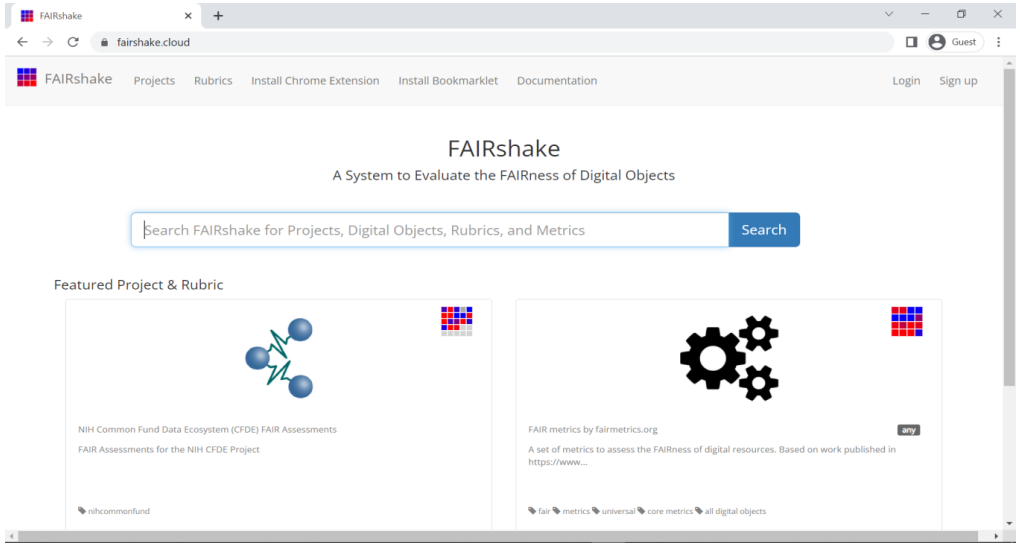
Soluții de evaluare

- F-UJI Automated FAIR Data Assessment Tool (FAIRsFAIR)
  - Implementează metricile DOMS (Data Object Assessment Metrics), pentru care sunt definite cerințe de implementare și metode de evaluare
  - Codul sursă este diponibil sub licență free in Github



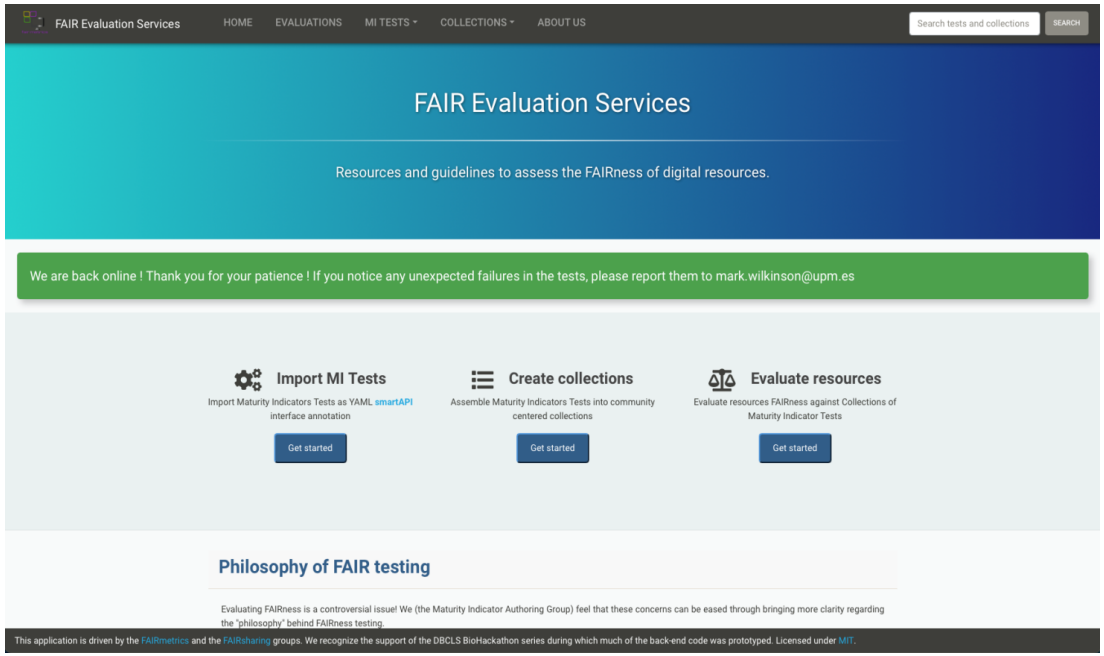
Soluții de evaluare

- FAIRshake - A System to Evaluate the FAIRness of Digital Objects
  - Descris in *FAIRshake: toolkit to evaluate the findability, accessibility, interoperability, and reusability of research digital resources*. DOI: https://doi.org/10.1101/657676
  - Permite definirea de metrici FAIR la nivel de comunitate științifică, cu facilități atașate de evaluare FAIR manuală, semiautomată sau automată
  - Exemplu de studiu de caz: NIH Common Fund Data Ecosystem (CFDE) FAIR Assessments



Soluții de evaluare

- Cadru pentru evaluarea automata a metricilor FAIR
  - Ecosistem de instrumente
  - FAIR Evaluator: https://fairsharing.github.io/FAIR-Evaluator-FrontEnd



Bibliografie

[1] Mons, Barend et al. Cloudy, Increasingly FAIR; Revisiting the FAIR Data Guiding Principles for the European Open Science Cloud. Information Services & Use, 37(1), pp. 49-56, 2017, DOI: 10.3233/ISU-170824.

[2] A. Jacobsen, R. Kaliyaperumal, L.O. Bonino da Silva Santos, B. Mons, E. Schultes, M. Roos & M. Thompson. A generic workflow for the data FAIRification process. Data Intelligence 2(2020), 56–65, DOI: 10.1162/dint_a_00028.

[3] Science Europe. Practical Guide to the International Alignment of Research Data Management. Extended Edition, January 2021, DOI: 10.5281/zenodo.4915862.