Despre evaluarea FAIR a datelor de cercetare

Eveniment de instruire pe tema
”Constituirea capacităților”

București, 28.01.2021

Ella Magdalena Ciupercă, Gabriel Neagu
ICI București

Conținut

- Principii și cerințe FAIR
- Cadrul general FAIR
- Metrici FAIRsFAIR pentru evaluarea obiectelor de date (scurt tutorial)
- RDA - Model de maturitate a datelor FAIR (sinteză)

Principiile FAIR

- ***F**indability, **A**ccessibility, **I**nteroperability* și ***R**eusability* reprezintă **un set minimal de principii și practici relaționate, dar independente și separabile**, care să permită atât calculatoarelor, cât și oamenilor să identifice, să acceseze, să interopereze și să reutilizeze datele și metadatele de cercetare
- Principiile FAIR **nu sunt reguli stricte**, ci concepte care inspiră eforturile de a ridica nivelul de calitate al datelor, **cu dezavantajul unor ambiguități și diferențe de interpretare**
- **Cerințele FAIR** reprezintă primul nivel de detaliere, cu caracter generic, a semnificației acestor principii

Cerințe FAIR

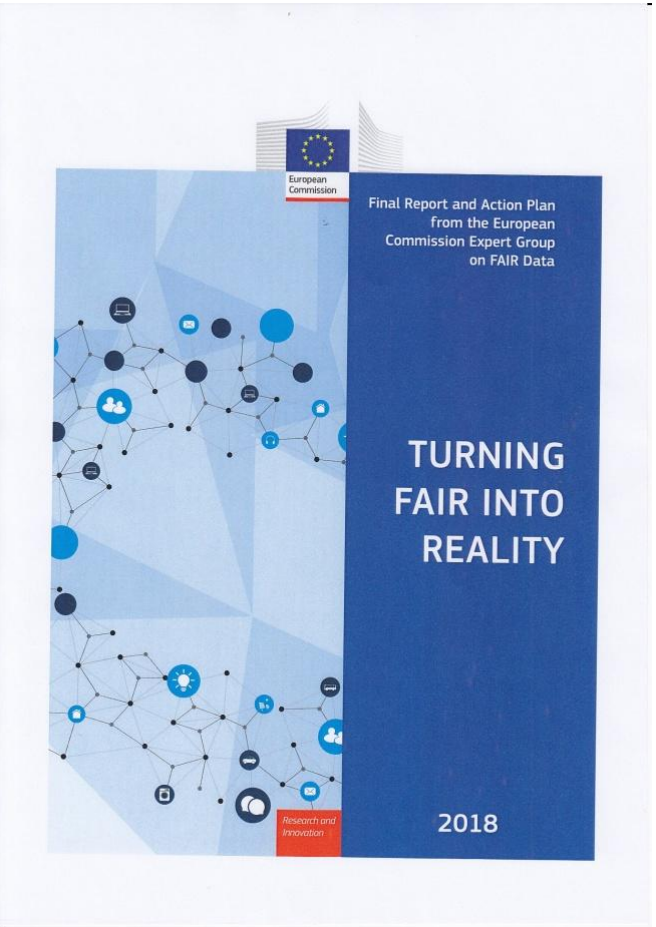
- **Date *Findable***:
  - F1. (meta)datele au asignat un identificator unic la nivel global și persistent
  - F2. datele sunt descrise prin intermediul metadatelor
  - F3. (meta)datele sunt înregistrate sau indexate pe un suport care permite căutarea
  - F4. Metadatele specifică clar și explicit identificatorul datelor pe care le descriu
- **Date *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
    - A1.2. este definită, după caz, o procedură de acces
  - A2. (meta)datele sunt accesibile chiar și după ce datele nu mai sunt disponibile

Cerințe FAIR

- **Date *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
- **Date *Reusable***:
  - R1. (meta)datele au o multitudine de atribute precise și relevante (ex. fișiere README):
    - R1.1. (meta)datele sunt eliberate cu o licență clară și accesibilă de utilizare (ex. Creative Common)
    - R1.2. (meta)datele sunt asociate cu proveniența lor
  - R1.3. (meta)datele respectă standardele comunității relevante din domeniu

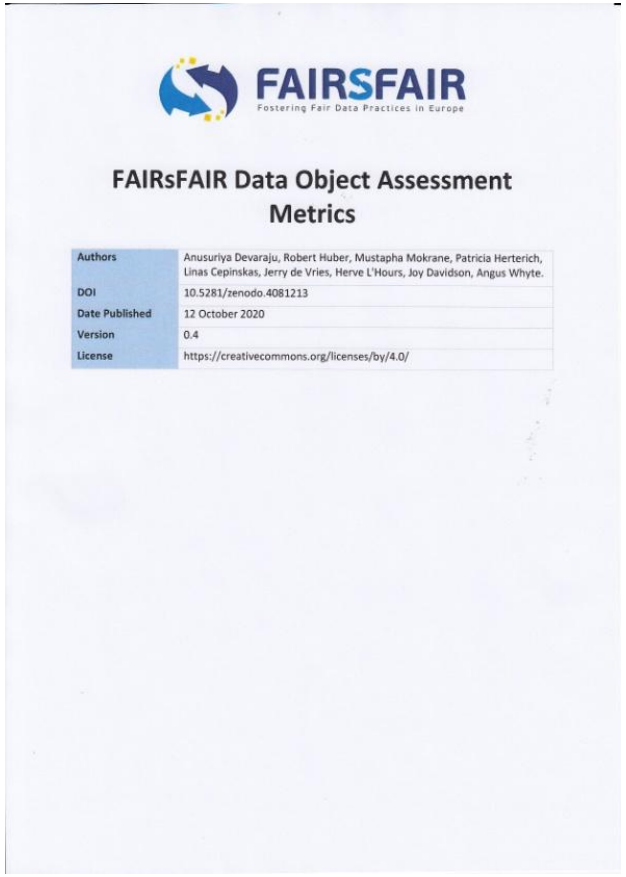
Cadrul general FAIR

- **Turning FAIR into Reality** – Final report and action plan from the EC Expert Group on FAIR Data (2018)
- Componentele "realității FAIR":
  - **Obiecte digitale FAIR**: date, software, alte resurse de cercetare
  - **Ecosistemul FAIR**: servicii, specificații metadate, depozite digitale, DPM
  - **Suport de interoperabilitate**: formate de date, standarde de metadate, instrumente, infrastructuri
  - **FAIR accesibil pentru om și calculator**: analiza și integrarea datelor în infrastructuri federative
  - **Competențe** în domeniile Data science și Data stewardship
  - **Metrici și indicatori** de implementare, corelați cu **stimulente**



Metrici FAIRsFAIR pentru evaluarea obiectelor de date

- **FAIRsFAIR Data Object Assessment Metrics** – DOAM (oct.2020)
  - se aplică subsetului de obiecte digitale reprezentat de **datele de cercetare**, colectate sau create în scopul analizei științifice
  - un obiect de tip dată de cercetare se referă la **date**, **metadate** și **documentație** (cum ar fi politici și proceduri)
  - pentru servicii implementate pe calculator (de ex. depozite digitale) și documente digitale, se au în vedere **teste automate** pentru metrici
    - unele aspecte (exactitate, relevanță, bogație, pluralism) specificate în principiile FAIR **necesită încă mediere și interpretare umană**
  - **17 metrici**: 5 – Findable / 4 – Accesible / 3 – Interoperable / 5 - Reusable


FAIRSFAIR
Fostering Fair Data Practices in Europe

FAIRsFAIR Data Object Assessment Metrics

Authors	Anusuriya Devaraju, Robert Huber, Mustapha Mokrane, Patricia Herterich, Linas Cepinskas, Jerry de Vries, Herve L'Hours, Joy Davidson, Angus Whyte.
DOI	10.5281/zenodo.4081213
Date Published	12 October 2020
Version	0.4
License	https://creativecommons.org/licenses/by/4.0/

1- Identificator unic global (F1)

- **Obiectului de tip data** i se atribuie un identificator unic global
  - Un identificator trebuie asociat cu un singur obiect de date
    - Un repozitor poate atribui un asemenea identificator la încărcarea datelor sau metadatelor
  - Exemple de **identificatori unici ai datelor**:
    - Internationalized Resource Identifier (IRI), Uniform Resource Identifier (URI), Digital Object Identifier (DOI)
  - **Evaluare**: se verifică dacă identificatorul este specificat pe baza unei scheme de identificare unice la nivel global
  - **Referințe suport**:
    - Uniform Resource Identifier (URI) Generic Syntax (RFC 3986), https://tools.ietf.org/html/rfc3986
    - Identificatori prelucrați de organizația FAIRsharing, https://fairsharing.org/standards/?q=&selected_facets=type_exact:identifier%20schema

2- Identificator persistent (F1)

- **Obiectului de tip data** i se atribuie un identificator persistent (PID)
  - Un URL de tip HTTP este o adresă unică la nivel global pentru o resursă web, dar poate să nu fie persistentă
  - Identificatorii bazați pe Handle System, Digital Object Identifier (DOI), Archival Resource Key (ARK) **sunt atât globali, cât și persistenți**
  - Asigurarea persistenței este o **responsabilitate comună** pentru un furnizor de servicii PID și pentru clienții săi (de exemplu, depozitele digitale de date)
  - **Evaluare**: se verifică dacă identificatorul este specificat pe baza unei scheme PID general acceptate, inclusiv existența ”landing page” și informația de acces la date în metadate
  - **Referințe suport**:
    - Generic PID definitions, Initial Persistent Identifier Policy for the EOSC, https://doi.org/10.5281/zenodo.3574202

3- Metadate descriptive de bază (F2)

- **Metadatele** includ elemente descriptive de bază (autor, titlu, identificator de date, editor, data publicării, rezumat, cuvinte cheie)
  - Metadatele de bază sunt **independente de domeniu** și **permit identificarea și localizarea datelor**, inclusiv citarea lor
  - Metadatele necesare se stabilesc pe baza **ghidurilor comune pentru citarea datelor** (de exemplu, DataCite, ESIP) și a **recomandărilor de metadate pentru descoperirea datelor**, de exemplu: EOSC Datasets Minimum Information (EDMI), DataCite Metadata Schema, W3C Recommendation Data on the Web Best Practices, Data Catalog Vocabulary
  - **Evaluare**: se accesează pagina de metadate (prin identificatorul datelor) unde se verifică existența metadatelor de bază
  - **Referințe suport** – standarde de metadate:
    - FAIRsharing, https://fairsharing.org/standards/
    - RDA Metadata Directory (General Research Data Standards), https://rd-alliance.github.io/metadata-directory/subjects/general.html

4- Includerea identificatorului datelor în metadate (F3)

- **Metadatele** includ identificatorul datelor pe care le descriu
  - Permite utilizatorilor să poată **descoperi și accesa datele prin metadate**
  - **Evaluare**: se accesează pagina de metadate (prin identificatorul datelor) unde se verifică existența unui identificator identic al datelor și dacă link-ul pentru accesarea datelor este funcțional
  - **Referințe suport**:
    - FAIR Signposting Profile, https://signposting.org/FAIR/

5- Metadate care pot fi regăsite (F4)

- **Metadatele** sunt furnizate într-o formă care permite **regăsirea automată**
  - Metadatele pot fi disponibile prin puncte finale multiple (endpoints)
    - dacă datele sunt găzduite într-un depozit de date, acesta poate să disemineze metadatele acestora printr-un protocol de recoltare a metadatelor (metadata harvesting protocol), ca de ex. Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)
  - Metadatele pot fi inserate ca date structurate pe o pagină de date pentru a fi **utilizate de către motoarele de căutare web** sau pentru a fi **disponibile ca date deschise conectate** (*open link data*)
  - **Evaluare**: se verifică accesul automat la metadate
    - existența și corectiudinea *endpoints*
    - structurarea corectă (*search engine friendly*) a informației din *landing page*
  - **Referințe suport**:
    - Documentație de referință Google privind reprezentarea datelor structurate de tip Dataset https://developers.google.com/search/docs/data-types/dataset

6- Informația de acces la date (A1)

- **Metadatele** conțin nivelul de acces și condițiile de acces la date
  - Atât nivelul de acces, cât și condițiile de acces sunt **informații necesare pentru a obține accesul potențial la date**
    - Ex: date deschise fără restricții de acces, date deschise accesibile de la o dată, date cu acces restricționat, acces doar la metadate.
  - Principiul „**cât de deschis posibil, cât de restricționat necesar**”
  - **Evaluare**: se verifică existența în metadate a informației de acces la date
  - **Referințe suport**:
    - Public domain licenses, https://creativecommons.org/share-your-work/public-domain
    - EU Vocabulary on access rights, https://op.europa.eu/en/web/eu-vocabularies/at-dataset/-/resource/dataset/access-right
    - Open Digital Rights Language (ODRL) Information Model 2.2, https://www.w3.org/TR/odrl-model/
    - Controlled Vocabulary for Access Rights, http://vocabularies.coar-repositories.org/documentation/access_rights/

7- Protocol de comunicații standard pentru metadate (A1)

- **Metadatele** sunt accesibile printr-un protocol de comunicații standard
  - Exemple sunt **protocoalele de nivel aplicție** : HTTP, HTTPS, FTP și AtomPub. Este de evitat diseminarea metadatelor folosind un protocol proprietar.
  - **Evaluare**: se verifică protocolul utilizat pentru accesarea paginii de preluare a cererii de acces la date (pagina de metadate)
  - **Referințe suport**:
    - Exemple de protocoale nivel aplicație, https://en.wikipedia.org/wiki/Application_layer
    - IANA Protocol Registries, https://www.iana.org/protocols

8- Protocol de comunicații standard pentru date (A1)

- **Datele** sunt accesibile printr-un protocol de comunicații standard
  - Exemple sunt **protocoalele de nivel aplicție** : HTTP, HTTPS, FTP, TFTP, SFTP, FTAM și AtomPub. Este de evitat diseminarea metadatelor folosind un protocol proprietar.
  - **Evaluare**: se verifică protocolul utilizat în identificatorul datelor
  - **Referințe suport**:
    - Exemple de protocoale nivel aplicație, https://en.wikipedia.org/wiki/Application_layer
    - IANA Protocol Registries, https://www.iana.org/protocols

9- Conservarea datelor (A2)

- **Metadatele** rămân disponibile, chiar dacă datele nu mai sunt disponibile
  - Accesul continuu la metadate depinde de practica de conservare a depozitului de date
  - **Evaluare**: se face la nivel de *repozitory,* nu la nivel de obiect de date
    - In cazul unui PID, o parte din metadate sunt conservate în registrul administrat de furnizorul PID
  - **Referințe suport**:
    - DMP Common Standards WG, https://www.rd-alliance.org/groups/dmp-common-standards-wg

10- Reprezentarea formală a metadatelor (I1)

- **Metadatele** sunt reprezentate folosind un limbaj formal de reprezentare a cunoștințelor
  - Exprimarea metadatelor unui obiect de date folosind o reprezentare formală a cunoștințelor permite calculatorului să le proceseze și să asigure diverse modalități de schimb de date
  - Exemple de limbaje de reprezentare a cunoștințelor sunt RDF, RDFS și OWL.
  - **Evaluare**: se verifică accesarea metadetelor în reprezentare formală
    - De ex. pentru reprezentările RDF se utilizează limbajul SPARQL
  - **Referințe suport**:
    - SPARQL Protocol for RDF, https://www.w3.org/TR/rdf-sparql-protocol/
    - Best Practice Recipes for Publishing RDF Vocabularies, https://www.w3.org/TR/swbp-vocab-pub/

11- Metadate cu resurse semantice (I1)

- **Metadatele** folosesc resurse semantice
  - Îmbogățirea metadatelor poate facilita căutarea datelor și interoperabilitatea datelor din diferite surse
  - **Ontologia, tezaurul** și **taxonomia** sunt tipuri de resurse semantice cu diferite grade de expresivitate și complexitate de calcul
  - **Evaluare**: se verifică conținutul metadatelor prin filtrarea numelor comune (rdf, rdfs, owl etc.) și compararea celor rămase cu intrări din ontologii, tezaure cunoscute.
  - **Referințe suport**:
    - Publishing and consuming Linked Data embedded in HTML, https://www.w3.org/2001/sw/interest/ldh/
    - Linked Open Vocabularies (LOV), https://lov.linkeddata.es/dataset/lov

12- Legături cu entități conexe (I3)

- **Metadatele** includ legături între date și entitățile sale conexe
  - Datele legate la entitățile sale conexe **sporesc potențialul de reutilizare** a acestora
  - Informațiile de legătură ar trebui să fie capturate **ca parte a metadatelor**
  - Un set de date poate fi legat de versiunea sa anterioară, de seturi de date sau de resurse conexe
  - **Se recomandă PID pentru entitățile conexe** (de exemplu, ORCID pentru contribuabili, DOI pentru publicații și ROR pentru instituții)
  - **Evaluare**: se verifică dacă metadatele indică legături între date și entități conexe și dacă aceste legături sunt active
  - **Referințe suport**:
    - DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.3, https://doi.org/10.14454/7xq3-zf69

13- Metadate pentru conținutul datelor (R1)

- **Metadatele** specifică conținutul datelor
  - **Exemple de proprietăți** care specifică conținutul datelor: tip resursă (date sau colecție de date), variabilă măsurată sau observată, metodă, format și dimensiune a datelor
  - **Standardele de metadate cu scop general** (Datacite Metadata Schema și Schema.org) oferă elemente pentru a reprezenta descrierile de conținut, dar varietatea structurilor de date și dimensiunile acestora pot face dificile testele de conformitate
  - **Evaluare**: se verifică prezența elementelor care reprezintă descrierea conținutului datelor în documentul de metadate, se accesează datele prin URL-ul indicat în metadate și se compară descrierea cu proprietățile reale ale datelor
  - **Referințe suport**:
    - CSV on the Web: A Primer, https://www.w3.org/TR/tabular-data-primer/
    - Apache Tika (an example of content analysis toolkit), https://tika.apache.org/

14- Licența de utilizare a datelor (R1.1)

- **Metadatele** includ informații despre licența în baza căreia datele pot fi refolosite
  - Se recomandă aplicarea licențelor indiferent dacă datele sunt publice, restricționate sau dedicate unor anumiți utilizatori
  - Licențele pot fi standard (Creative Commons, Open Data Commons Open Database Licence) sau licențe personalizate și declarații de drepturi de proprietate care specifică condițiile în care datele pot fi refolosite
  - **Evaluare**: se verifică existența în metadate a informațiilor despre licență (nume sau URI) și posibilitatea de a obține info suplimentare din surse externe
  - **Referințe suport**:
    - SPDX license registry, https://spdx.org/licenses/
    - Open Digital Rights Language (ODRL), https://www.w3.org/TR/odrl-model/
    - Creative Commons Rights Expression Language, https://creativecommons.org/ns

15- Proveniența datelor (R1.2)

- **Metadatele** includ informații despre proveniența sau generarea datelor
  - **Proveniența** (descendența) datelor reprezintă istoricul unui set de date, inclusiv persoanele, entitățile și procesele implicate în crearea sau colectarea, versionarea, publicarea, gestionarea și întreținerea lor pe termen lung
  - Pentru contributor se poate utiliza PID (ORCID pentru cercetători, DOI pentru date) sau link-uri la înregistrări de proveniență (PROV sau VoID-Vocabulary of Interlinked Datasets)
  - **Evaluare**: se verifică existența în metadate a informațiilor minimale despre proveniență
  - **Referințe suport**:
    - PROV Model Primer, https://www.w3.org/TR/prov-primer/
    - W3C Recommendation Data on the Web Best Practices (8.4 Data Provenance), https://www.w3.org/TR/dwbp/#metadata
    - PAV- Provenance, Authoring and Versioning ontology: https://pav-ontology.github.io/pav/

16- Standardul comunității referitor la metadate (R1.3)

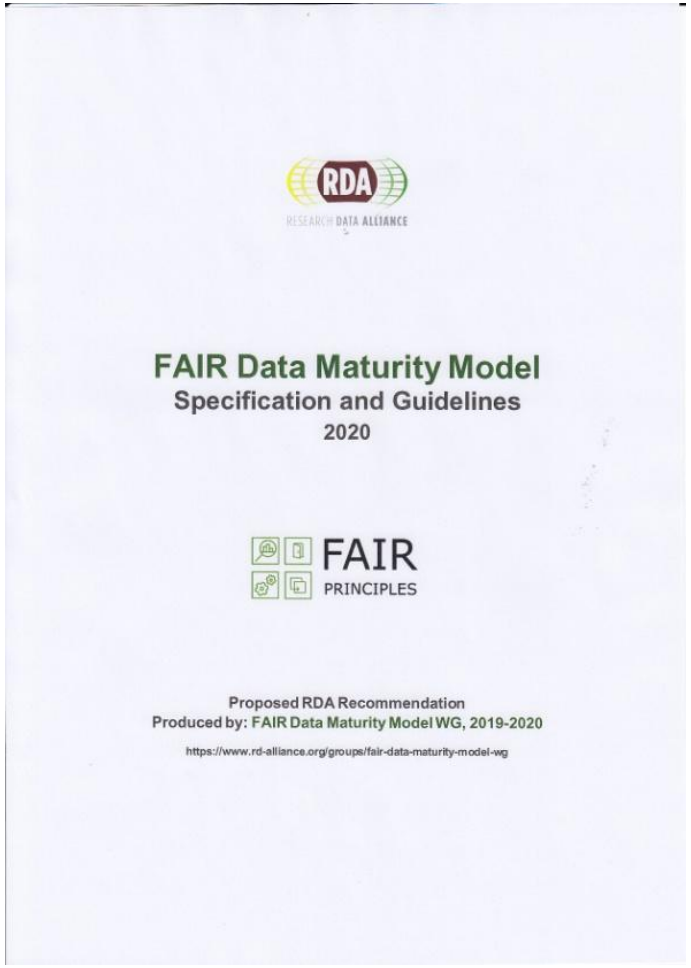
- **Metadatele** respectă un standard recomandat de comunitatea de cercetare țintă
  - În plus față de metadatele de bază necesare pentru descoperirea datelor (metrica 3), metadatele care asigură reutilizarea datelor trebuie să fie conforme cu standardele de metadate aprobate de comunitate
  - Exemple: geospațiu: ISO19115; biodiversitate: Darwin Core, ABCD; științe sociale: DDI; astronomie: International Virtual Observatory Alliance TS
  - **Evaluare**: se identifică standardele de metadate orientate pe domeniu care sunt utilizate în depozitul digital de date (prin raportare la un catalog specializat) și se testează conformitatea metadatelor asociate unui identificator de date cu unul din aceste standarde
  - **Referințe suport**:
    - RDA Metadata Standards Catalog, https://rdamsc.bath.ac.uk/
    - FAIRsharing, https://fairsharing.org/standards/

17- Formatul fișierului de date (R1.3)

- **Datele** sunt disponibile într-un format de fișier recomandat de comunitatea de cercetare țintă
  - Datele ar trebui puse la dispoziție într-un format de fișier care este agreat de comunitatea de cercetare, pentru a permite partajarea și reutilizarea
  - Aceste formate ar trebui să fie adecvate și pentru stocarea / arhivarea pe termen lung
  - **Evaluare**: se extrag din metadate informații despre despre formatul fișierului (mime-type) și se verifică dacă este un format de fișier deschis și de termen lung
  - **Referințe suport:**
    - Liste de formate de fișiere științifice de uz general sau specifice:
      - http://justsolve.archiveteam.org/index.php/Scientific_Data_formats
      - https://en.wikipedia.org/wiki/List_of_file_formats#Scientific_data_(data_exchange)
    - Exemple de formate de fișiere recomandate pe baza tipurilor de date: https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats.aspx

Model de maturitate a datelor FAIR

- **RDA FAIR Data Maturity Model** – Specification and Guidelines (iunie 2020)
  - Propune criterii de evaluare a nivelului de implementare a principiilor FAIR
- Componentele modelului:
  - **Indicatori** – aspecte individuale specifice evaluării calității FAIR a datelor
  - **Priorități** - importanța relativă a indicatorilor
  - **Metoda de evaluare** - stabilirea valorilor indicatorilor pe baza rezultatelor evaluării



Indicatori DMM

- Răspund la întrebarea ”***Ce trebuie măsurat*** *pentru a evalua nivelul FAIR al unui obiect digital*” ?
- Repartizarea celor 41 de indicatori pe principiile FAIR
  - 7 *Findable*: 2 pentru date și 5 pentru metadate
  - 12 *Accesible*: 6 pentru date și 6 pentru metadate
  - 12 *Interoperable*: 5 pentru date și 7 pentru metadate
  - 10 *Reusable*: 2 pentru date și 8 pentru metadate
- Comparație DMMI / DOAM [%]:
  - Pondere metadate: 63 / 76
  - Repartizare pe principii FAIR: F - 17/**29**; A – **29**/24; I – **29**/18; R – 14/**29**

Priorități DMM

- Sunt definite trei niveluri de prioritate:
  - **Esențial**: indicatorul se referă la un aspect determinant pentru nivelul FAIR
  - **Important**: indicatorul abordează un aspect care ar crește considerabil nivelul FAIR
  - **Util**: aspectul abordat este dezirabil, dar nu indispensabil

	F	A	I	R	Total
Esențial	7	8	0	5	**20**
Important	0	3	7	4	**14**
Util	0	1	5	1	**7**
Total	**7**	**12**	**12**	**10**	**41**

Metoda de evaluare DMM

- **Pentru evaluare indicator** – două abordări:
  - **Măsurarea progresului**: accent pe gradul în care o resursă evaluată îndeplinește cerințele exprimate în indicator
    - **Nivelurile de maturitate** ale unui indicator:

      0 - nu se aplică

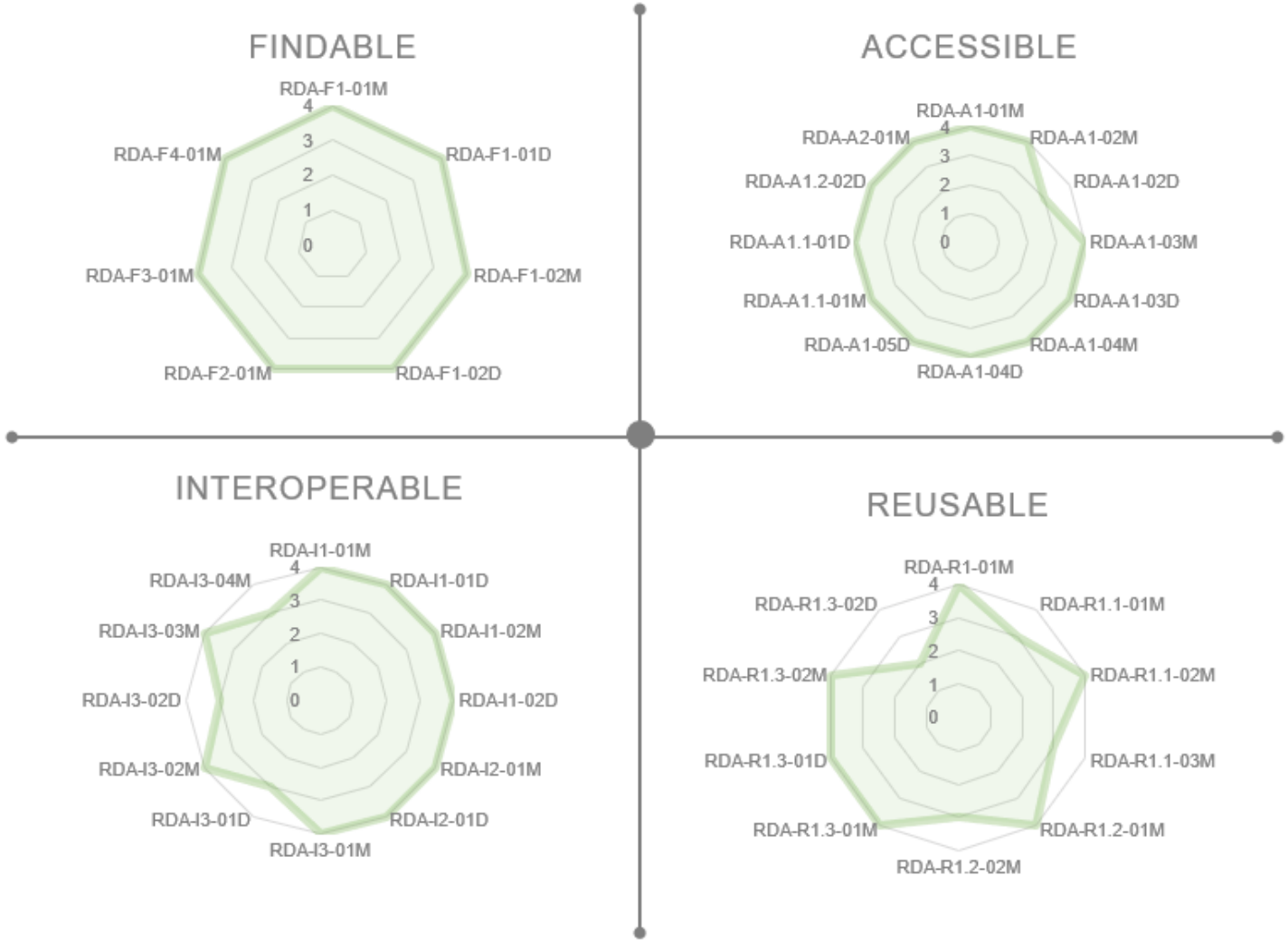
      1 - nu este luat în considerare

      2 - în curs de analiză sau în fază de planificare

      3 - în faza de implementare

      4 - complet implementat
  - **Măsurarea succesului sau a eșecului**: determină dacă o resursă evaluată reușește sau nu să atingă un anumit nivel țintă pentru un indicator
    - Este o abordare mai exigentă – succesul este consemnat numai **pentru atingerea nivelului 4 de maturitate**

Metoda de evaluare DMM



Sursa:

RDA FAIR Maturity Method

Metoda de evaluare DMM

- **Pentru evaluare resursă** (pe ansamblul indicatorilor):
  - Nivel 0 – nu este FAIR
  - Nivel 1 - FAIR indicatorii esențiali
  - Nivel 2 - FAIR indicatorii esențiali + 50% indicatori importanți
  - Nivel 3 - FAIR indicatorii esențiali + 100% indicatori importanți
  - Nivel 4 - FAIR indicatorii esențiali + 100% indicatori importanți + 50% indicatori utili
  - Nivel 5 - FAIR indicatorii esențiali + 100% indicatori importanți + 100% indicatori utili

Vă mulțumim pentru atenție !