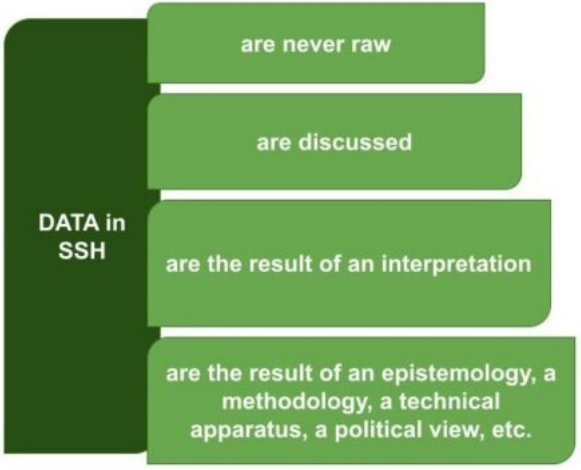
FAIR adatkezelés

Holl András – MTA KIK – HRDA

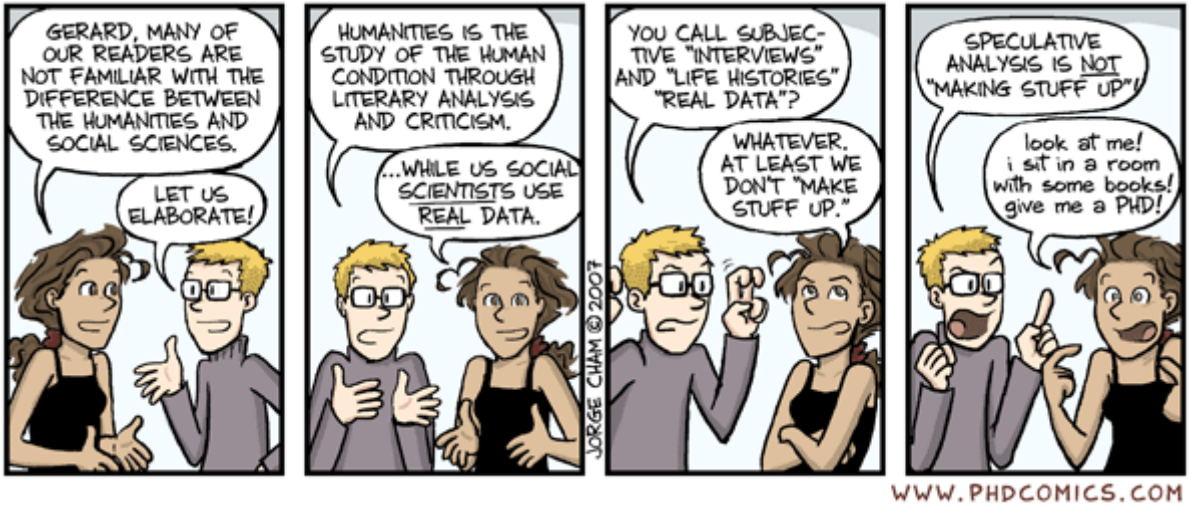
Módszeresen
TK Szociológiai Intézet, TÁRKI Adatbank
2021 Dec. 9.

Adatkoncepció a társadalomtudományokban



Karla Avanço

„First of all, it must be said that humanities researchers tend to be ambivalent about the concept of 'data' and that "[t]here are issues surrounding […] the acceptance of the 'research data concept'". In short, they just don't use the word "data", but talk about "sources", "research materials" etc., which leads to the fact that the whole "data talk" doesn't appeal to them." Ulrike Wuttke, „Here be dragons"



PHD Comics
https://phdcomics.com/comics/archive.php?comicid=908

Kutatási adatok – Open Research Data / FAIR Research Data

A nyílt hozzáférésű tudomány (Open Science) mozgalom eredeti célkitűzései szerint a tudományos közlemények szabad hozzáférhetőségéhez hasonlóan a kutatási adatok is szabadon hozzáférhetővé kellett váljanak. Ma a releváns adatkezelési paradigma a FAIR.

„A kutatási adatok legyenek annyira nyíltak, amennyire lehetséges, és annyira zártak, amennyire szükséges."

„As open as possible, as closed as necessary."

Áthelyezett hangsúly a társadalomtudományokban:

adatok/formátumok *helyett* módszerek/kockázatok

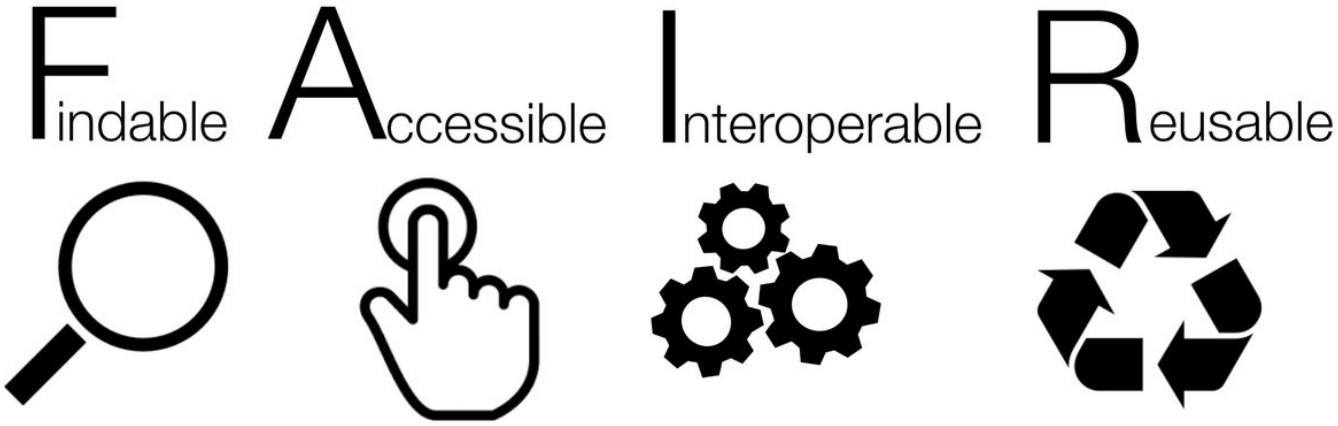
FAIR – Findable, Accessible, Interoperable, Reusable (megtalálható, hozzáférhető, szabványos, újrafelhasználható)

2016-ban jelentek meg a FAIR alapelvek.

- Wilkinson, Mark D. et al. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific data* vol. 3 160018. 15 Mar. 2016, doi:10.1038/sdata.2016.18

- FORCE11 – The FAIR data principles
https://www.force11.org/group/fairgroup/fairprinciples

- GoFAIR – FAIR Principles
https://www.go-fair.org/fair-principles/



Rec. 3: A model for FAIR Data Objects
Implementing FAIR requires a model for FAIR Data Objects which by definition have a PID linked to different types of essential metadata, including provenance and licencing. The use of community standards and sharing of code is also fundamental for interoperability and reuse.

FAIR

- F: Findable (megtalálható)

- Egyedi azonosítóval rendelkezik (pl. DOI);

- gazdagon el van látva leíró adatokkal (metaadatok);

- be van jegyezve, indexálva van valamilyen nyilvános, kereshető szolgáltatásba.

F**A**IR

- A: Accessible (hozzáférhető)

- Az egyedi azonosítón keresztül mind a leíró adatok, mind maguk az adatok elérhetőek valamilyen szabványos protokoll segítségével;

- a protokoll nyílt, ingyenes, elterjedt;

- a protokoll lehetővé teszi az azonosítást és a jogosultságkezelést, amennyiben erre szükség van;

- a leíró adatok akkor is hozzáférhetőek, ha az adatok már nem.

FA**I**R

- I: Interoperable (szabványos)

- Az adatok és a leíró adatok szabványos és értelmezhető formában vannak;

- a leíráshoz, adatrögzítéshez, dokumentációhoz használt szótárak maguk is eleget tesznek a FAIR alapelveknek;

- sok kereszthivatkozás segíti az értelmezést.

FAI**R**

– R: Reusable (újrafelhasználható)

- **Egyértelmű felhasználási licenc áll rendelkezésre;**
- **az adatok származása, keletkezése jól dokumentált;**
- **a tudományterületi szabványoknak megfelel.**

The FAIR Guiding Principles	
Findable:	
F1	Data and metadata are assigned a globally unique and persistent identifier
F2	Data are described with rich metadata (defined by R1 below)
F3	Metadata clearly and explicitly include the identifier of the data it describes
F4	Data and metadata are registered or indexed in a searchable resource
Accessible:	
A1	Data and metadata are retrievable by their identifier using a standardized communications protocol
A1.1	The protocol is open, free, and universally implementable
A1.2	The protocol allows for an authentication and authorization procedure, where necessary
A2	Metadata are accessible, even when the data are no longer available
Interoperable:	
I1	Data and metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2	Data and metadata use vocabularies that follow FAIR principles
I3	Data and metadata include qualified references to other (meta)data
Reusable:	
R1	Data and metadata are richly described with a plurality of accurate and relevant attributes
R1.1	Data and metdata are released with a clear and accessible data usage license
R1.2	Data and metadata are associated with detailed provenance
R1.3	Data and metadata meet domain-relevant community standards

(Wilkinson et al. 2016)

FAIR principle	FAIR Maturity Indicator
F1	Data and metadata identifiers are unique and persistent
F2	Metadata are structured (weak) or grounded in shared vocabularies (strong)
F3	Data and metadata identifiers are included explicitly in metadata
F4	Searchable in web-based search engines
A1.1	Uses open free protocol for data & metadata retrieval
A1.2	Data and metadata authentication and authorization
A2	Metadata persistence
I1	Data and metadata knowledge representation language (weak or strong)
I2	Metadata uses FAIR vocabularies or ontologies (weak or strong)
I3	Metadata contains qualified outward references
R1.1	Metadata includes a license for data usage (weak or strong)
R1.2	Metadata includes provenance (weak)
R1.3	Metadata contains community standards (weak)

weak - readable only by humans; strong - readable by machines

(Wilkinson et al. 2019)

A FAIR alapelvek maradnak, értelmezésük, részletes kidolgozásuk nem zárult le.

RDA FAIR Data Maturity Model WG – Indicators

F	F1	F1-01M	Metadata is identified by a persistent identifier	Recommended
	F1	F1-02M	Metadata is identified by a universally unique identifier	Recommended
	F1	F1-01D	Data is identified by a persistent identifier	Mandatory
	F1	F1-02D	Data is identified by a universally unique identifier	Mandatory
	F2	F2-01M	Sufficient metadata is provided to allow discovery, following domain/discipline-specific metadata standard	Recommended
	F2	F2-02M	Metadata is provided for the discovery-related elements defined by the RDA Metadata IG, as much as possible and relevant, if no domain/discipline-specific metadata standard is available	Recommended
	F3	F3-01M	Metadata includes the identifier for the data	Mandatory
	F4	F4-01M	Metadata or landing page is harvested by general search engine	Recommended
	F4	F4-02M	Metadata is harvested by or submitted to domain/discipline-specific portal	Recommended
	F4	F4-03M	Metadata is indexed in institutional repository	Recommended

	A1	A1-01M	Metadata includes information about access conditions	Optional
	A1	A1-01D	Data is available for manual download	Recommended
	A1	A1-02D	Data is available for automatic download	Recommended
	A1	A1-02M	Metadata identifier resolves to a metadata record	Optional
	A1	A1-03D	Data identifier resolves to a data file	Mandatory
	A1	A1-03M	Metadata is accessed through standardised protocol	Recommended
	A1	A1-04D	Data is accessible through standardised protocol	Recommended
A	A1.1	A1.1-01D	Data is accessible through a free access protocol	Mandatory
	A1.1	A1.1-02D	Data is accessible through an open-source access protocol	Recommended
	A1.1	A1.1-01M	Metadata is accessible through a free access protocol	Mandatory
	A1.1	A1.1-02M	Metadata is accessible through an open-source access protocol	Recommended
	A1.2	A1.2-01D	Data is accessible through an access protocol that supports authentication	Recommended
	A1.2	A1.2-02D	Data is accessible through an access protocol that supports authorisation	Recommended
	A1.2	A1.2-01M	Metadata includes information relevant for access control	Mandatory
	A2	A2-01M	Metadata is guaranteed to remain available after data is no longer available	Mandatory

I	I1	I1-01M	Metadata uses knowledge representation expressed in standardised format	Recommended
	I1	I1-02M	Metadata uses machine-understandable knowledge representation	Optional
	I1	I1-03M	Metadata uses self-describing knowledge representation	Optional
	I1	I1-01D	Data uses knowledge representation expressed in standardised format	Recommended
	I1	I1-02D	Data uses machine-understandable knowledge representation	Optional
	I1	I1-03D	Data uses self-describing knowledge representation	Optional
	I2	I2-01M	Metadata uses standard vocabularies	Recommended
	I2	I2-01D	Data uses standard vocabularies	Recommended
	I2	I2-02M	Metadata uses FAIR-compliant vocabularies	Optional
	I2	I2-02D	Data uses FAIR-compliant vocabularies	Optional
	I3	I3-01M	Metadata includes references to other metadata	Recommended
	I3	I3-01D	Data includes references to other data	Recommended
	I3	I3-02M	Metadata includes sufficiently qualified references to other metadata	Recommended
	I3	I3-02D	Data includes sufficiently qualified references to other data	Optional

R	R1	R1-01M	Sufficient metadata is provided to allow reuse, following domain/discipline-specific metadata standard	Recommended
	R1	R1-02M	Metadata is provided for the reuse-related elements defined by the RDA Metadata IG, as much as possible and relevant, if no domain/discipline-specific metadata standard is available	Recommended
	R1.1	R1.1-01M	Metadata includes information about the licence under which the data can be reused	Mandatory
	R1.1	R1.1-02M	Metadata refers to a standard reuse licence	Recommended
	R1.1	R1.1-03M	Metadata includes licence information in the appropriate element of the metadata standard used	Mandatory
	R1.1	R1.1-04M	Metadata refers to a machine-understandable reuse licence	Optional
	R1.1	R1.1-06M	Metadata includes information about consent for reuse (e.g. for personal data)	Recommended
	R1.2	R1.2-01M	Metadata includes provenance information according to community-specific guidelines	Recommended
	R1.2	R1.2-02M	Metadata includes provenance information according to a cross-domain language	Optional
	R1.3	R1.3-01M	Metadata complies with a community standard	Mandatory
	R1.3	R1.3-01D	Data complies with a community standard	Mandatory
	R1.3	R1.3-02M	Metadata is expressed in compliance with a machine-understandable community standard	Optional
	R1.3	R1.3-02D	Data is expressed in compliance with a machine-understandable community standard	Optional

Más szempontrendszerek is léteznek: a TRUST és a CARE

TRUST alapelvek:
Transparency, Responsibility, User Focus, Sustainability and Technology
Lin, D., Crabtree, J., Dillo, I. et al.
The TRUST Principles for digital repositories. Sci Data 7, 144 (2020).
https://doi.org/10.1038/s41597-020-0486-7

CARE alapelvek:
Collective Benefit, Authority to Control, Responsibility, Ethics
(benszülött közösségek adatai)
https://www.gida-global.org/care

FAIR – mi a kutatók feladata?

A FAIR alapelveknek való megfelelés részben az adatokat előállító (gyűjtő, kompiláló, megfigyelő, stb.) kutatón múlik, részben az adatokat megőrző és hozzáférhetővé tevő repozitóriumon. A kutató feladatai:

- gazdag metaadatok (F)
- repozitóriumi elhelyezés egyedi állandó azonosítóval (repozitórium választás) (F)

- az adatok és a metaadatok szabványos formában vannak (szótár, tezaurusz, ontológia választás) (I)
- kereszthivatkozások segítik az értelmezést (I)

- megfelelő felhasználási licenc (licenc választás) (R)
- az adatok jól dokumentáltak (R)
- megfelelnek a tudományos közösség szabványainak (R)

FAIR a kutatók szempontjából

Miért szükséges a FAIR alapelvek alkalmazása?

- **Külső kényszer: pályázatok, folyóiratok**
- **Jobb kutatási gyakorlat**
- **Átláthatóbb kutatás, jobban reprodukálható eredmények**
- **GDPR**
- **Több adat (különböző kutatásokból) ?**

FAIR is a journey

Karel Luyben: 20 év múlva a kutatási adatok 50%-a …

FAIR szintek:

- **eredmények ellenőrizhetősége más kutatók számára**
- **adatok újrafelhasználása tudományterületen belül**
- **adatok újrafelhasználása más területen (?)**
- **adatok gépi értelmezése**

FAIR veszélyek – digitális szakadék (?)

További információk, irodalom

SSH szószedetek, tezauruszok
Controlled vocabularies, UK Data Archive
https://www.data-archive.ac.uk/managing-data/standards-and-procedures/controlled-vocabularies/
pl. INDIGENOUS POPULATIONS

The CLARIN & SSHOC Vocabulary Initiative
https://www.clarin.eu/blog/clarin-sshoc-vocabulary-initiative

FAIR szószedetek
https://fairvocabularies.github.io/about

OPERAS White Paper Common Standards and FAIR Principles
https://doi.org/10.5281/zenodo.5653414

Open Data for Humanists, A Pragmatic Guide
https://dh.tcd.ie/dh/open-data-for-humanists-a-pragmatic-guide/

FAIR Data in Social Sciences and Humanities: a Triple Open Science Training Session
https://roadtofair.hypotheses.org/327

“Here be dragons”: Open Access to Research Data in the Humanities
https://dhmethods.hypotheses.org/262

White Paper on implementing the FAIR principles for data in the Social, Behavioural, and Economic Sciences. RatSWD Working Paper 274/2020
http://doi.org/10.17620/02671.60

INTERNATIONAL CLASSIFICATION OF CRIME FOR STATISTICAL PURPOSES (ICCS)
https://www.unodc.org/unodc/en/data-and-analysis/statistics/iccs.html

Françoise, Genova; Jan, Magnus Aronsen; Oya, Beyan; Natalie, Harrower; András, Holl; Rob, W.W. Hooft; Pedro, Principe; Ana, Slavec et al.: Recommendations on FAIR metrics for EOSC; Luxembourg, Luxemburg : Publications Office of the European Union (2021)
http://real.mtak.hu/119439/

Cost-Benefit analysis for FAIR research data – Cost of not having FAIR research data
European Commission, Luxembourg: Publications Office of the European Union, 2018
ISBN 978-92-79-98886-8
http://doi.org/10.2777/02999

Köszönöm a figyelmet!
Beszéljük meg!