







Quality requirements for open data in biomedicine hindrances and emerging standards René Bernard, Dept. for Exp. Neurology 03/16/2018

CHARITÉS

RANKENHAUS

Open Data

- Openness of data- a quality standard vital part of PREMIER
- Needed for transparency, reproducibility and re-use
- Data sharing many hindrances
- Non-existing standards within many fields of biomedical research



Mega-Repositories







zenodo

- Undocumented ,,data dumps"
- Limited quality control, consistency problems
- Often not linked to original research paper
- How to re-analyse ?

Official Gene Symbol	Entrez Gene Description (Homo sapiens)	itrez Gene Description (Homo sapiens) Excel Date Conversi				
DEC1	deleted in esophageal cancer 1	1-Dec				
MARCL	mitochondrial amidoxime reducing component 1	1-Mar				
MARCH1	membrane associated ring finger 1					
MARCH2	membrane associated ring finger 2					
MARC2	mitochondrial amidoxime reducing component 2	2-Mar				
MARCH3	membrane associated ring finger 3 3-Mar					
MARCH4	membrane associated ring finger 4	4-Mar				
MARCH5	membrane associated ring finger 5	5-Mar				
MARCH6	membrane associated ring finger 6	6-Mar				
MARCH7	membrane associated ring finger 7	7-Mar				
MARCH8	membrane associated ring finger 8	8-Mar				
MARCH9	membrane associated ring finger 9	9-Mar				
MARCH10	membrane associated ring finger 10	10-Mar				
MARCH11	membrane associated ring finger 11	11-Mar				
SEPTI	septin 1	1-Sep				
SEPT2	septin 2	2-Sep				
SEP.T3	septin 3	3-Sep				
SEPT4	septin 4	4-Sep				
SEPTS	septin 5	5-Sep				
SEPT6	septin 6	6-Sep				
SEPT7	septin 7	7-Sep				
SEPT8	septin 8	8-Sep				
SEPT9	septin 9	9-Sep				
SEPT10	septin 10	10-Sep				
SEPT11	septin 11	11-Sep				
SEPT12	septin 12	12-Sep				
SEPT14	septin 14	14-Sep				
SEPT15	15 kDa selenoprotein	15-Sep				

https://library.medicine.yale.edu/blog/do-not-let-excel-deplete-your-gene-list

FAIR data principles

 Image: Specific product of Research Communications and e-Scholarship
 Findable
 Accessible
 Interoperable
 Reuse

 Image: Specific product of Research Communications and e-Scholarship
 Image: Scholarship
 Image: Scholarship

- Set of guiding principles
- Each point has 3-4 quality levels
- First established 2015
- Data FAIRness embraced by governments and adopted funding bodies (EC, NIH, G20)



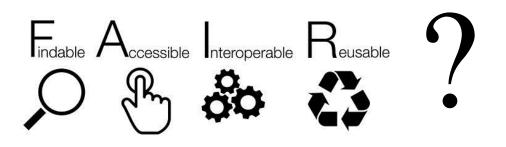
FAIR is NOT a standard

- Definition of goals
- Do not define **how** to achieve FAIRness of research data
- FAIR is **not** equal to open (accessible under well-defined conditions)
- Wide (mis)interpretation of data FAIRness

General issues hindering data re-use in Life Science

- Too many identifiers for the same concept
- Free text descriptions of methodologies
- No standard symboles or variables for measurements
- No one size –fits all approach



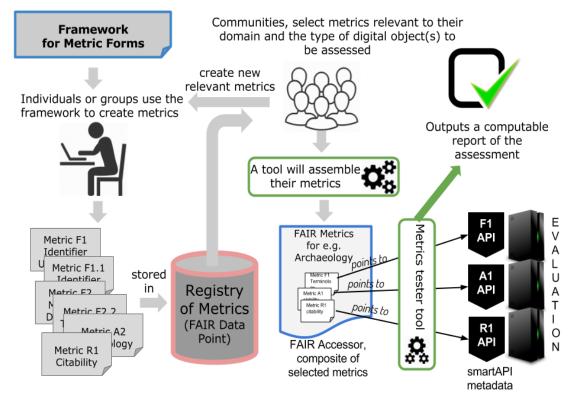


How to achieve data FAIRness in biomedical research ?



Top Down: Frameworks for RDM

Framework for the development and execution of the FAIR Metrics



SCIENCE EUROPE



Presenting a Framework for Discipline-specific Research Data Management JANUARY 2018

www.scienceeurope.org/

http://fairmetrics.org/

How did clinicians solve the standard problem?

- Common Data Elements (CDEs)
 - Are standardized key terms or concepts
 - Developed by the NIH for clinical research and patient registries
 - To improve data quality & comparison from multiple studies and electronic health records across sites and time

Name of CDE	Definition	Query / Instructions	Provenance	Value Set	Resource Link
Societal Sex	Text designations that identify gender.	Self- identified gender.	This value set is from Health Level Seven International Table 0001	Ambiguous; Female; Male; Not applicable; Other; Unknown; Male-to-female transsexual; Female-to-male transsexual	

• further subdivided: Core, Supplemental - Highly Recommended, Supplemental, or Exploratory

CDEs – organized in repository & specialty sites

- Permits harmonization across diverse areas; linking to other existing standards and terminologies (cancer, rare diseases, low back pain..)
- Subfields developed disease-specific CDEs

			▼ CDEs		Tools	:	▼ Learn	
	NIH CDE Repository	NEW Myalgic Encephalomyelitis/Chronic Fatigue Syndrome and Biomechanical Devices in TBI Recommendations Are you using the NINDS CDEs? Citing CDE Usage Let us know through our Feedback Form Learn to Use the CDEs: Study Start-Up Tutorial						
(stroke volume	Q. Search	Streamline	Your	Neuroscien	ce Clinical	CDEs Now Available Review	CD Devel
Browse by	Stroke volume measurement	🛓 Export Search 🔻	Research us	ing conte	nt standards that	enable clinical	General (CDEs that cross dis	
	Stroke Adjudication Worksheet - stroke TIA symptoms type		investigators to s	investigators to systematically collect, analyze, and share data			Amyotrophic Lateral Scleros	· · ·
HRQ —		NEI	across the research community.				Cerebral Palsy NEW	
Jency for H Julity Stroke Adjudication Worksheet - stroke TLA sub type Stroke Adjudication Worksheet - stroke TLA other symptoms type		National tye institute	National Lye Institute The NINDS strongly encourages researchers who receive funding fr Institute to ensure their data collection is compatible with these co elements (CDEs). Learn more about the CDE Project.					
						ith these common data	Epilepsy	
	Stroke Adjudication Worksheet - neuroimaging multiple CT stroke finding indicator Questionnaire for Verifying Stroke Free Status (QVSFS) - stroke status				-		Friedreich's Ataxia	
DA —	Ouestionnaire for Verifying Stroke Free Status (OVSFS) - stroke result	NLM National Library of Medicine	Launch Your Studies Faste		Incorporate CDEs Into Systems	Learn About the CDE Project	Headache (Version 2.0) NEW	/
inal Instit	sing Stroke Adjudication Worksheet - stroke TIA cardioembolic source symptoms type	National Library of Medicine	Case report form modules	1	Search for current CDEs	Project overview and background	Huntington's Disease	
			Standardized da		Download CDE	Meetings and	Mitochondrial Disease	
			element definition	ons	metadata Download Case	Presentations Collaboration with	Multiple Sclerosis	

https://commondataelements.ninds.nih.gov

Requirements for preclinical CDEs

- Useful for standardization of how experimental details and procedures are reported
- Wide adoption within a community and across others
- Accepted as key variables in field-specific databases
- Level of translatability (correlation with human CDEs)

Preclinical Common Data Elements

...shall not further standardize test procedures



But standardize the necessary descriptions

Preclinical Common Data Elements in TBI



HOME ABOUT DATA HOW TO POLICY NEWS



PRECLINICAL COMMON DATA ELEMENTS

Home » Data » Preclinical CDEs

Preclinical Traumatic Brain Injury Common Data Elements

Public Review Period Round 2: 12/01/17-02/28/18 (December 2017-February 2018)

Preclinical TBI CDE Round 2 PUBLIC REVIEW ZIP FILE 🗐

- TBI-Preclinical Working Group
- 40-50 researchers organized into 3 sub-working groups
 - General Health /Affective Disturbance (Depression/Anxiety/Social Interaction)
 - Cognition and Motor (Learning/Memory/Sensory/Motor)
 - Large Animal Models (Behavior)
- Subclassification similar to CDE
- Support: NINDS CDE Team

Organization of pCDE



Representation	Presentation		Data Type reco	mmended to use	Abbreviati	on for Variable Name		
Term			with a represe					
Anatomic Site			Alphanumeric AnatSite					
Category	the descriptive identification representing a level of in	tensity, defined	Alphanumeric		Cat			
	meaning, or subjective measurement							
Code	the selection from a system of defined categories for representation , of data, often defined using stratification or hierarchical organization;				Code			
Count	the quantity of the specified item		Numeric Values	;	Ct			
Date			Date or Date & Time Date					
Date/Time	the date and time when an event was observed or occ distinct from "Date" as there is a time element captur element.		Date or Date &	Time	DateTime			
Dose	the quantity of an agent (such as drug, substance or er administered, taken, or absorbed at one time		Numeric Values		Dose			_
Duration	the value measuring a quantity or period of time durin event or observation occurs			Variable Name		2	nder Test	Datatype
Frequency	the number of occurrences counted for an event with	Test subject in		LightDarkCycleT		Light/dark cycle type		Alphanumeric
	period	Test subject in		Acclimatization		Acclimatization to te		Numeric Values
Grade	the position on a scale of intensity or amount or quali	Equipment inf			ManufactNa Test apparatus manufacturer name ModelName Test apparatus model name		Alphanumeric Alphanumeric	
	to histology:				CTEquipCompartHeightMe			
		Equipment		erzquipeenipu		cymae rest equip	inent compartment neight	indification of the second sec
			info CTEquipCompartD		rtDiamtMeeCylinder Test - equip		pment compartment diam	Numeric Values
		Equipment info		CTEquipCompartLengthMe		le Cylinder Test - equipment compartment leng		Numeric Values
The pCDE sets can be		Equipment inf	ient info CTEquipC		tMirrorInd			Alphanumeric
		Room environ	ment	RoomIIIuminat	ionLevelVa	Room Illumination le	evel value	Numeric Values
ι	used as the building	Testing condit	ions	TestStartZeigTir	ne	Start of test in Zeitge		Date or Date & Time
1	blocks for study data collection forms.		Testing conditions		ie	End of test in Zeitgeb	er time	Date or Date & Time
Ĺ			rs	TrialDurationVa	al	Trial duration value		Numeric Values
<i>с</i>			ers	TrialTotalNum	- ckC of the second	Trials total number		Numeric Values
C		Software and software and software and software and software and software and software software and software so		*		Video motion trackin Video motion trackin		Alphanumeric Alphanumeric
		Software and		TestScoringMet		Test scoring method		Alphanumeric
		Data collected		CTForelimbCont			nb contacts with cylinder o	

LimbTyp

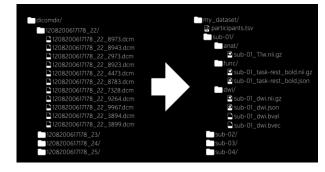
Limb type

Data collected

Alphanumeric

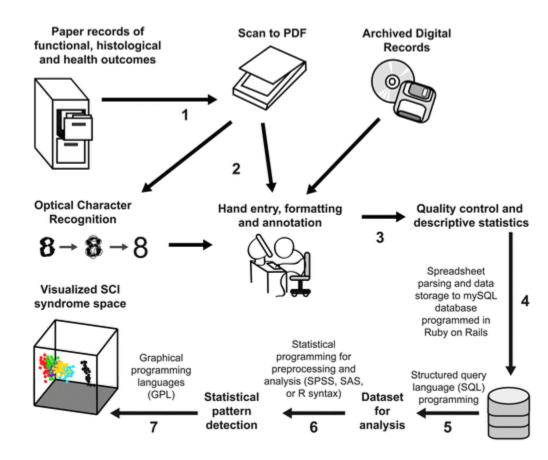
Example for community consensus -Neuroimaging

- Problem: no consensus in data organization
- Developed and published a standard Brain Image Data Structure (BIDS)
- naming remains consistent across all datasets
- Freely available data format converters ensured wide and fast adoption of BIDS
- Discipline-specific databases that drive on BIDS





Example for community consensus – Rodent Spinal Cord Injury Research



- Data archeology
- Reduce file drawer effect
- 4000 animals; 2700 variables; 13 labs
- Correlation pCDE-CDE
- VISION-SCI



Adam Ferguson, UCSF

Nielsen et al. Development of a Database for Translational SCI Research, *J Neurotrauma*, 2014 https://www.nature.com/articles/laban.1405

Example for community consensus - Rodent **Spinal Cord Injury Research**

• Latest development: Prospective SCI-Data plattform (scicrunch.org/odc-sci)



Nopen Data Commons for Spinal Cord Injury



Join the ODC-SCI Community to Accelerate SCI Research

Through the ODC portal you can learn about our diverse field, research methods, and stay informed on published research from the SCI community. With the growth of this community we hope to expand the amount of data sharing to promote transparency and reproducibility to promote our common goal to find a cure for SCI.



ABOUT ~



How to create your own open data community?



CREATE YOUR OWN COMMUNITY

Communities allow researchers to share and customize data from over 200 data resources.



Create or browse communities to explore personalized data portals for you or your group to work with Browse Resources Join the largest scientific resource registry and add, share, and search for new resources with your community. Search through Data Search across more than 200 data repositories

- Uploaded data are integrated in platforms framework
- Goal: "PubMed for data sources" + analytic capabilities
- Still needed: community buy-in

Summary

- Open Data shift in culture
- Value of preclinical data needs to mirror clincial
- Bottom up: databases with communities-established pCDEs
- Top down: Enduring support and funding structure
- Positive community examples shall lead the way



• Thank you very much!