CHARITÉ
UNIVERSITÄTSMEDIZIN BERLIN

PREMIER

Quality requirements for open data in biomedicine - hindrances and emerging standards

René Bernard, Dept. for Exp. Neurology

03/16/2018

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

figshare DRYAD MENDELEY DATA Put your research data online today zenodo

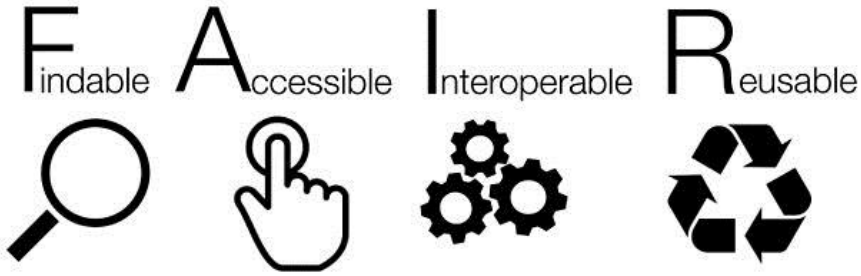
- Undocumented „data dumps“
- Limited quality control, consistency problems
- Often not linked to original research paper
- How to re-analyse ?

Official Gene Symbol	Entrez Gene Description (Homo sapiens)	Excel Date Conversion
DEC1	deleted in esophageal cancer 1	1-Dec
MARC1	mitochondrial amidoxime reducing component 1	1-Mar
MARCH1	membrane associated ring finger 1	
MARCH2	membrane associated ring finger 2	2-Mar
MARC2	mitochondrial amidoxime reducing component 2	
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
SEPT1	septin 1	1-Sep
SEPT2	septin 2	2-Sep
SEPT3	septin 3	3-Sep
SEPT4	septin 4	4-Sep
SEPT5	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

FORCE11
The Future of Research Communications and e-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)

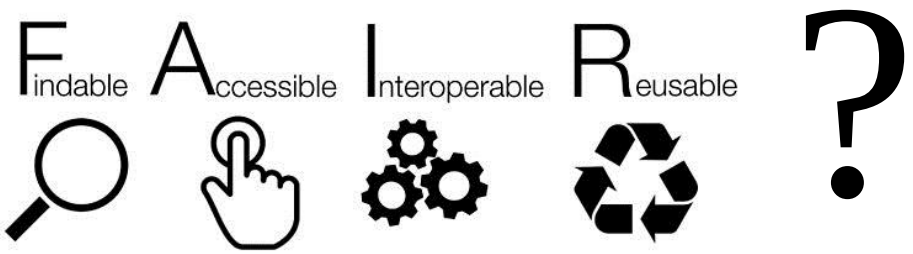
EOSC
The European Open Science Cloud

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

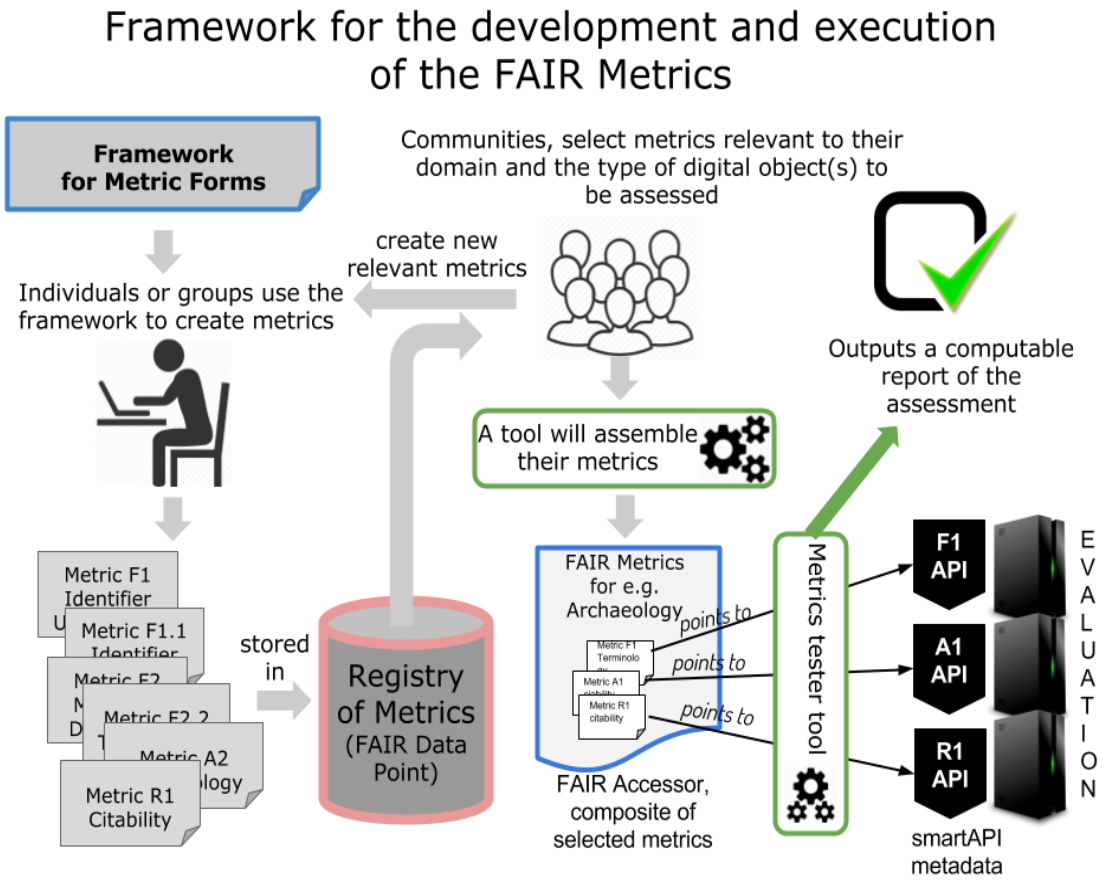
General issues hindering data re-use in Life Science

- 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



http://fairmetrics.org/

SCIENCE EUROPE

Science Europe Guidance Document

Presenting a Framework for Discipline-specific Research Data Management

JANUARY 2018

www.scienceeurope.org/

How did clinicians solve the standard problem?

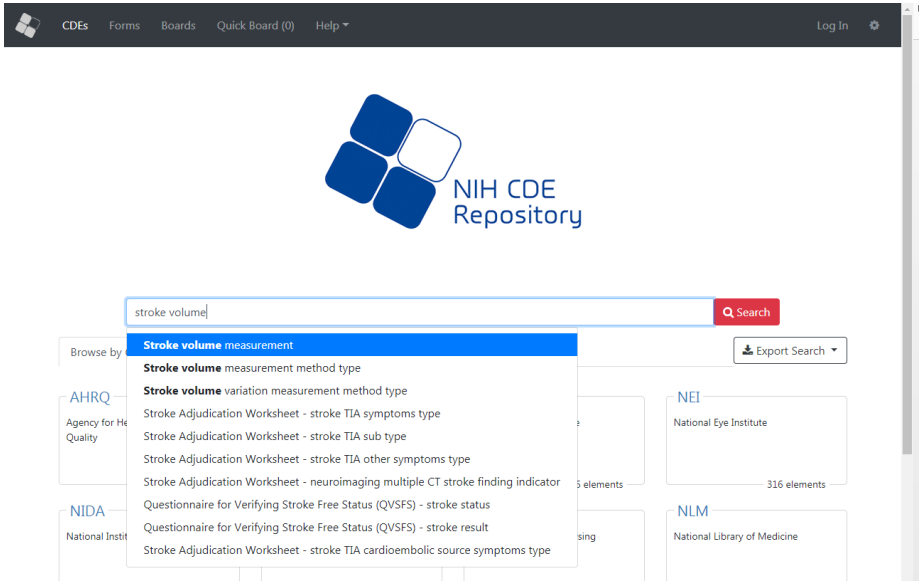
- Common Data Elements (CDEs)
  - Are standardized key terms or concepts
  - Developed by the NIH for clinical research and patient registries
  - To improve data quality & comparison from multiple studies and electronic health records across sites and time

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	GRDR

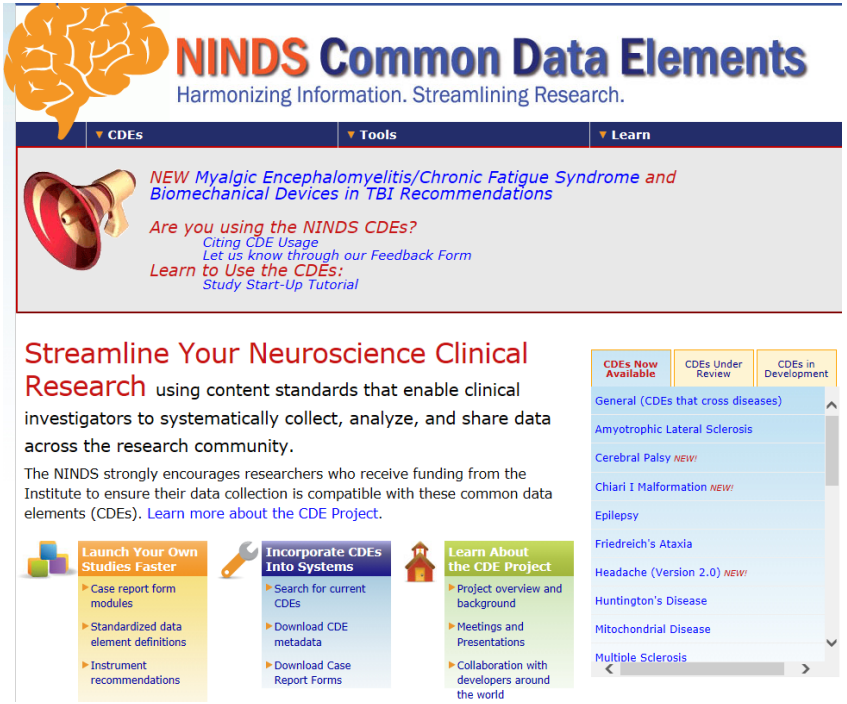
- further subdivided: Core, Supplemental - Highly Recommended, Supplemental, or Exploratory

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



https://cde.nlm.nih.gov



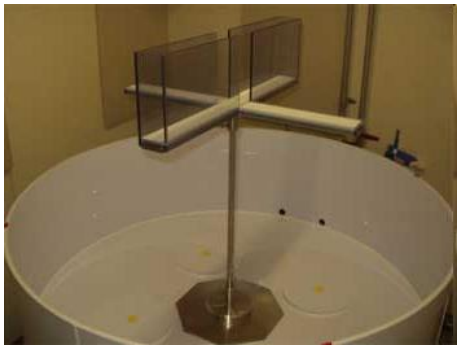
https://commondataelements.ninds.nih.gov

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)

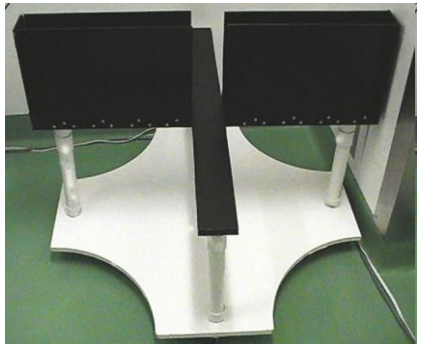
Preclinical Common Data Elements

…shall not further standardize test procedures



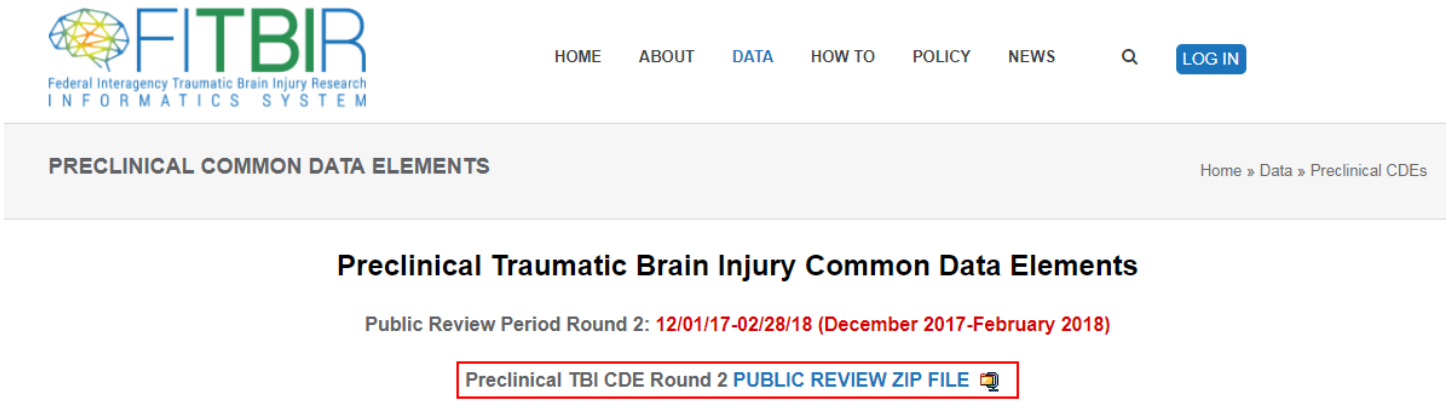






But standardize the necessary descriptions

Preclinical Common Data Elements in TBI



- 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

Presentation Terms

Representation Term	Definition	Data Type recommended to use with a representation term	Abbreviation for Variable Name
Anatomic Site	the named location of, or within, the body of a living being	Alphanumeric	AnatSite
Category	the descriptive identification representing a level of intensity, defined meaning, or subjective measurement	Alphanumeric	Cat
Code	the selection from a system of defined categories for representation of data, often defined using stratification or hierarchical organization;	Alphanumeric or Numeric Values	Code
Count	the quantity of the specified item	Numeric Values	Ct
Date	the date on which an event was observed or occurred	Date or Date & Time	Date
Date/Time	the date and time when an event was observed or occurred. This is distinct from "Date" as there is a time element captured in the data element.	Date or Date & Time	DateTime
Dose	the quantity of an agent (such as drug, substance or energy) 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		
Frequency	the number of occurrences counted for an event with period		
Grade	the position on a scale of intensity or amount or quali to histology;		

Cylinder Test

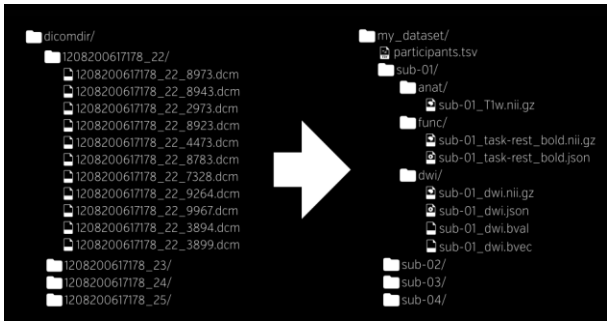
Category/Group	Variable Name	Title	Datatype
Test subject info	LightDarkCycleTyp	Light/dark cycle type	Alphanumeric
Test subject info	AcclimatizationTimeDur	Acclimatization to test room time duration	Numeric Values
Equipment info	TestApparatusManufactNa	Test apparatus manufacturer name	Alphanumeric
Equipment info	TestApparatusModelName	Test apparatus model name	Alphanumeric
Equipment info	CTEquipCompartHeightMe	Cylinder Test - equipment compartment height	Numeric Values
Equipment info	CTEquipCompartDiamtMea	Cylinder Test - equipment compartment diamet	Numeric Values
Equipment info	CTEquipCompartLengthMe	Cylinder Test - equipment compartment length	Numeric Values
Equipment info	CTEquipCompartMirrorInd	Cylinder Test - equipment mirror indicator	Alphanumeric
Room environment	RoomIlluminationLevelVa	Room Illumination level value	Numeric Values
Testing conditions	TestStartZeigTime	Start of test in Zeitgeber time	Date or Date & Time
Testing conditions	TestEndZeigTime	End of test in Zeitgeber time	Date or Date & Time
Test parameters	TrialDurationVal	Trial duration value	Numeric Values
Test parameters	TrialTotalNum	Trials total number	Numeric Values
Software and scoring	VideoMotionTrackSoftware	Video motion tracking software name	Alphanumeric
Software and scoring	VideoMotionTrackSoftware	Video motion tracking software version numbe	Alphanumeric
Software and scoring	TestScoringMethodTyp	Test scoring method type	Alphanumeric
Data collected	CTForelimbContactsCt	Cylinder Test - forelimb contacts with cylinder c	Numeric Values
Data collected	LimbTyp	Limb type	Alphanumeric

The pCDE sets can be used as the building blocks for study data collection forms.

Example for community consensus - Neuroimaging

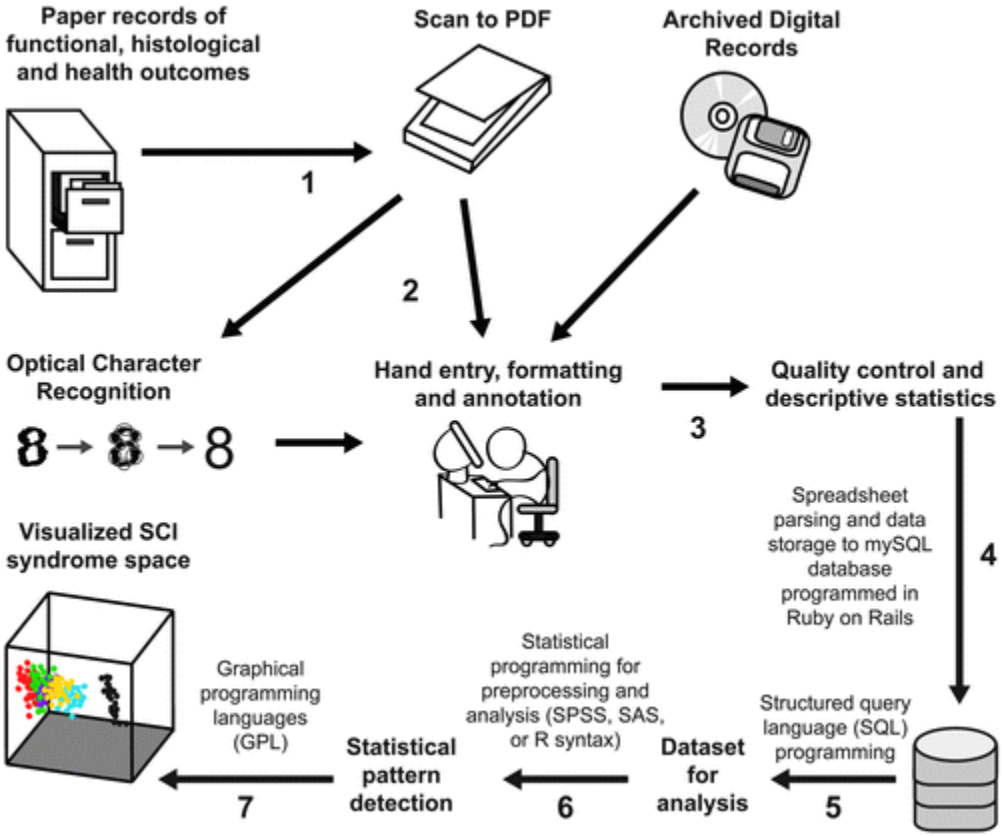
- **Problem:** no consensus in data organization
- Developed and published a standard - **B**rain **I**mage **D**ata **S**tructure (BIDS)
- naming remains consistent across all datasets
- Freely available data format converters

  ensured wide and fast adoption of BIDS
- Discipline-specific databases that drive on BIDS



http://bids.neuroimaging.io/

Example for community consensus – Rodent Spinal Cord Injury Research



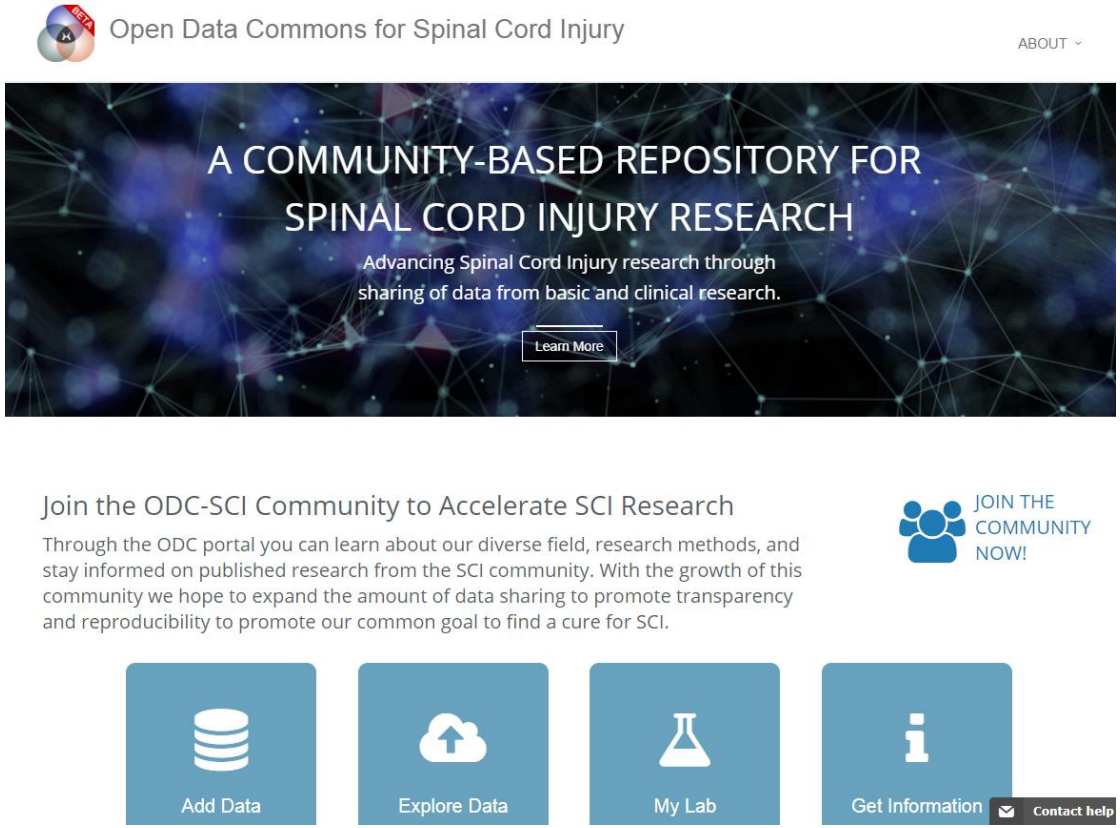
- Data archeology
- Reduce file drawer effect
- 4000 animals; 2700 variables; 13 labs
- Correlation pCDE-CDE
- VISION-SCI

Adam Ferguson, UCSF

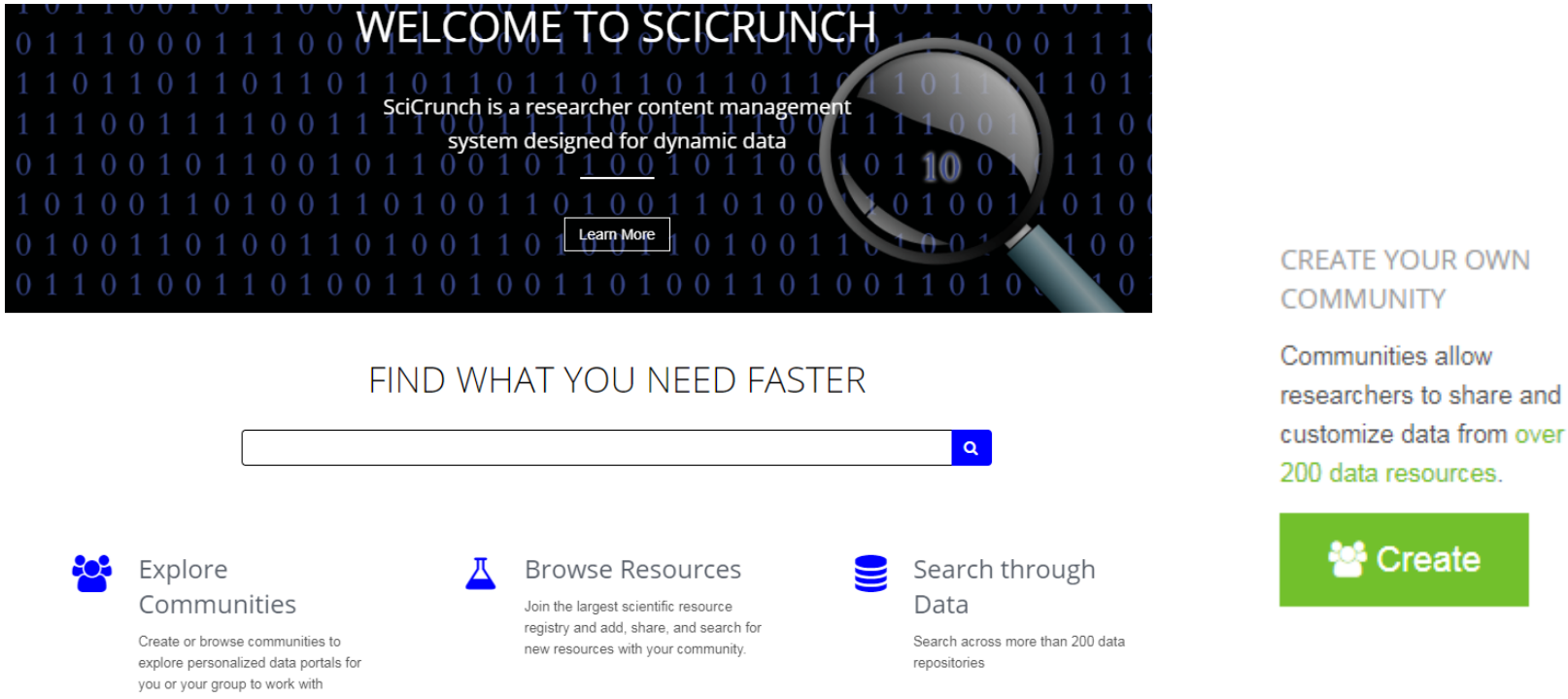
Nielsen et al. Development of a Database for Translational SCI Research, *J Neurotrauma*, 2014
https://www.nature.com/articles/laban.1405

Example for community consensus - Rodent Spinal Cord Injury Research

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



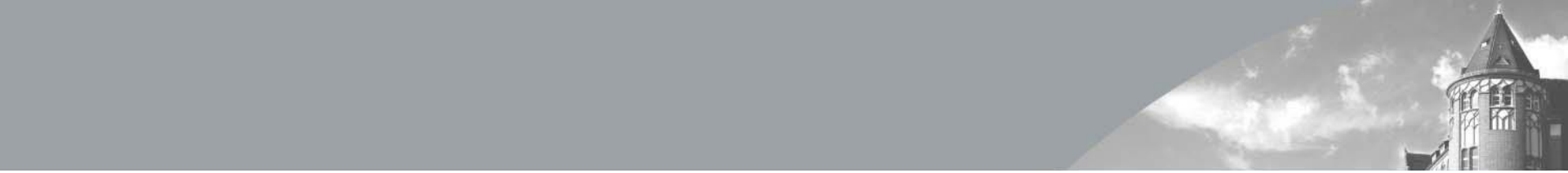
How to create your own open data community?



- 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!