LUXEMBOURG INSTITUTE OF **HEALTH**
RESEARCH DEDICATED TO LIFE

28
May 2019
Tuesday

Patients, trust and ethics in information privacy in e-Health – From fair to fair health

SPEAKER: Pr Jan-Eric LITTON (Sweden)

Supported by:

Fonds National de la Recherche Luxembourg

Partners:

uni.lu UNIVERSITÉ DU LUXEMBOURG

LISER LUXEMBOURG INSTITUTE OF SOCIO-ECONOMIC RESEARCH

LUXEMBOURG INSTITUTE OF SCIENCE AND TECHNOLOGY | LIST

CHL Centre Hospitalier de Luxembourg

CENTRE HOSPITALIER EMILE MAYRISCH

ITTM Information Technology for Translational Medicine

Patient, trust and ethics in information privacy in e-health, from FAIR to FAIR - Health principals

Lecture Series 28 May 2019, Luxembourg Institute of Health

Professor Jan-Eric Litton

Karolinska Institutet

Stockholm, Sweden.

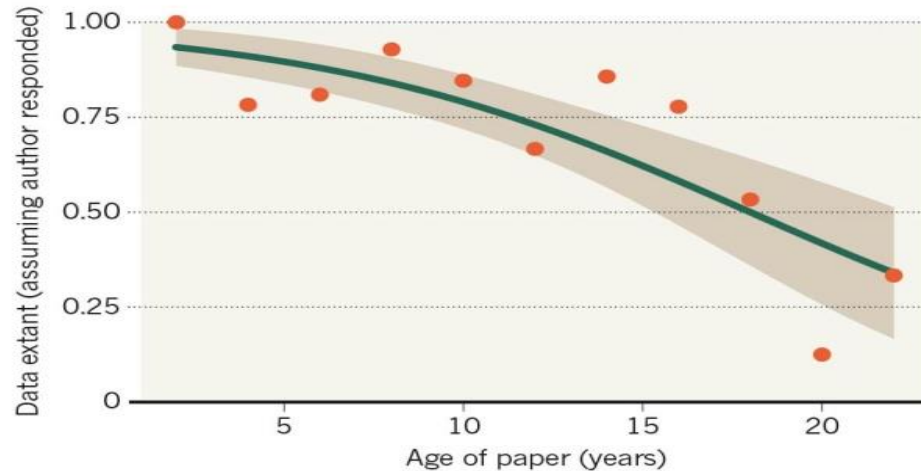# Why is FAIR-Health principals in e-health?

What is the problem??

Data loss is real and significant, while data growth is staggering

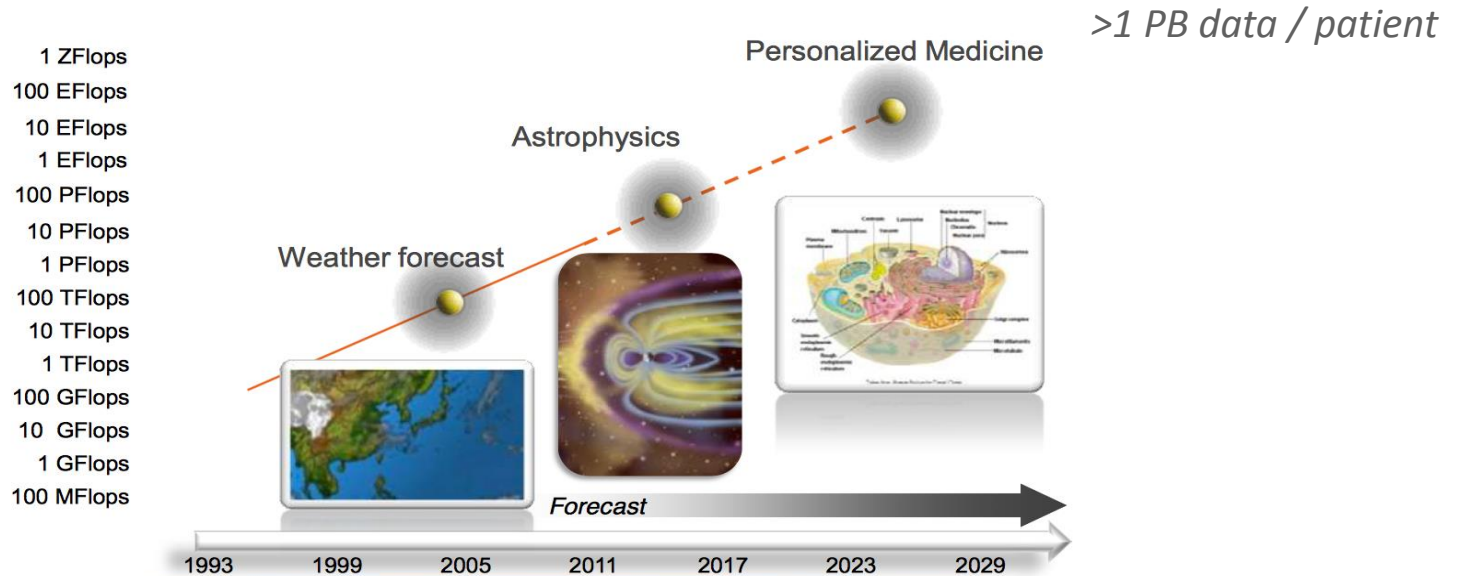# Data loss is real and significant, while data growth is staggering

MISSING DATA

As research articles age, the odds of their raw data being extant drop dramatically.

Nature news, 19 December 2013

Data loss is real and significant, while data growth is staggering

>1 PB data / patient

Personalized Medicine a key driver for Exascale ICT

from Intel 2012

Data loss is real and significant, while data growth is staggering

The amount of data we produce every day is truly mind-boggling. There are 2.5 quintillion (10^{18}) bytes of data created each day at our current pace, but that pace is accelerating with the growth of Internet of Things (IoT).

21 May 2018, Bernard Marr

Year	Global Internet Traffic
1992	100 GB per day
1997	100 GB per hour
2002	100 GB per second
2007	2,000 GB per second
2016	26,600 GB per second
2021	105,800 GB per second

Source: Cisco VNI, 2017.

One size does not fit all

|		
ANTI-DEPRESSANTS SSRIs	38%	
ASTHMA DRUGS	40%	
DIABETES DRUGS	43%	
ARTHRITIS DRUGS	50%	
ALZHEIMER'S DRUGS	70%	
CANCER DRUGS	75%	

Data loss is real and significant, while data growth is staggering - IoT
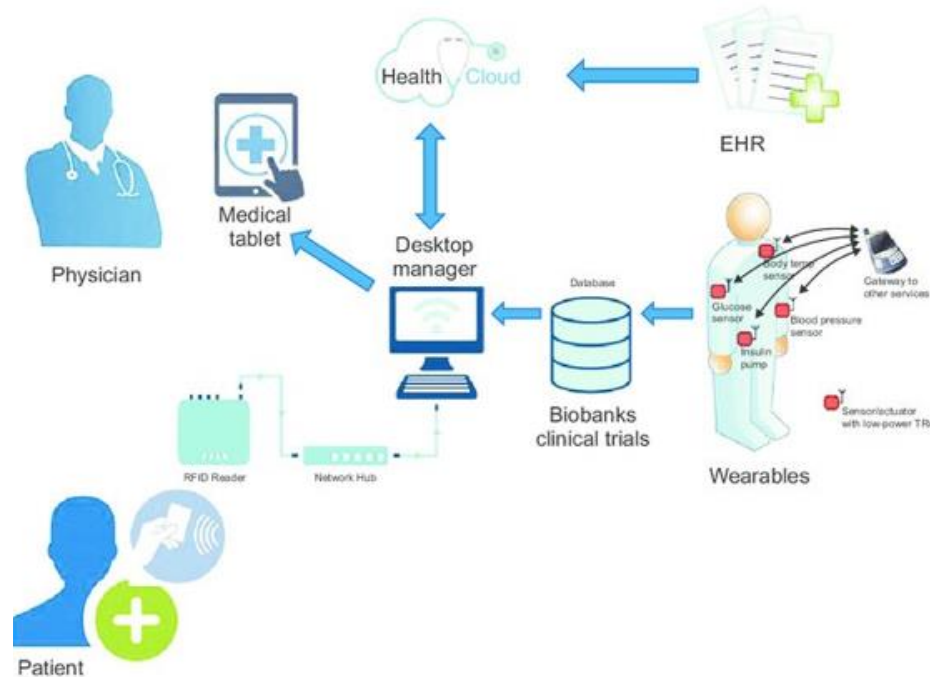
An illustration of how this revolution in medicine will look in a typical IoT in hospital

MedReD Example

o Using MedReD mobile application for studying how Schizophrenia patients enrolled at care centers nationwide benefit from physical excercise, collecting eQuestionnare and Accelerometer data

o Answer general quesionnaires about life habits, quality of life etc

o Go to gym and work-out.

o Answer questionnaire prior to work out session

o Turn on Accelerometer in mobile during excercise
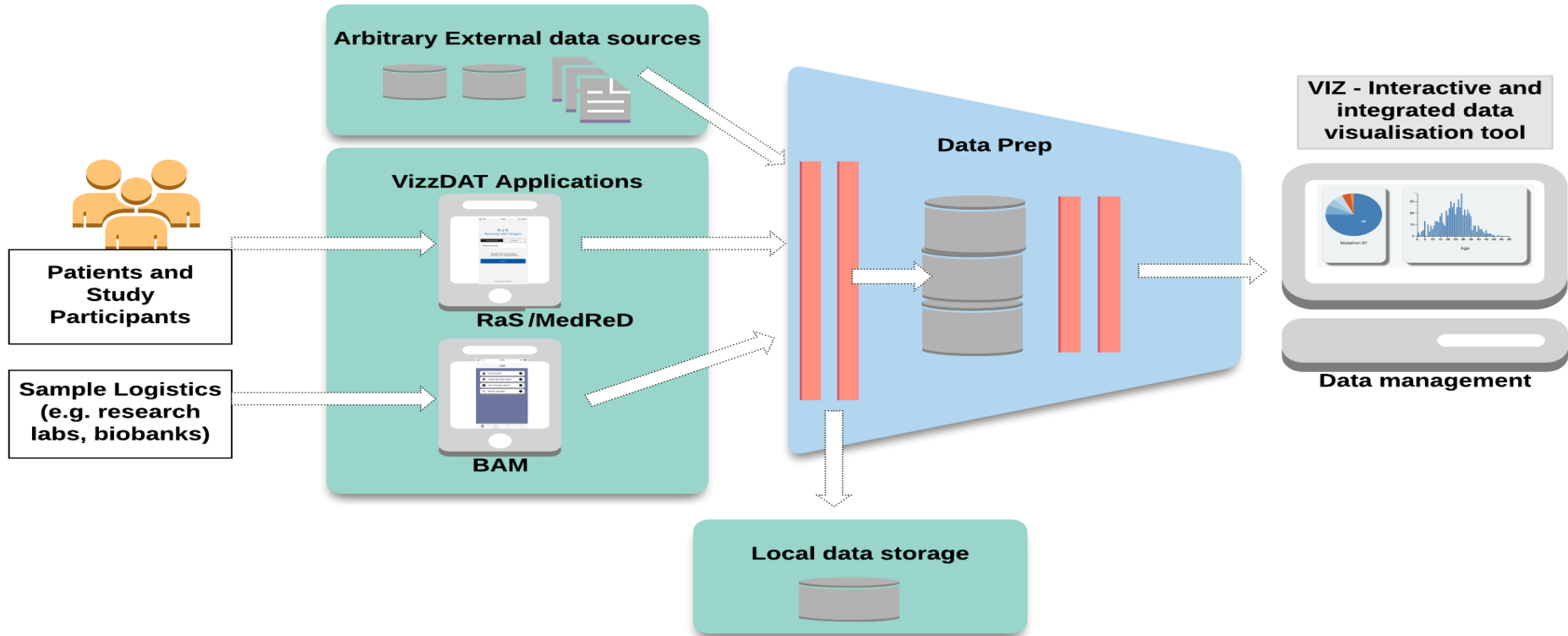
o Answer quesionnaire post work-out
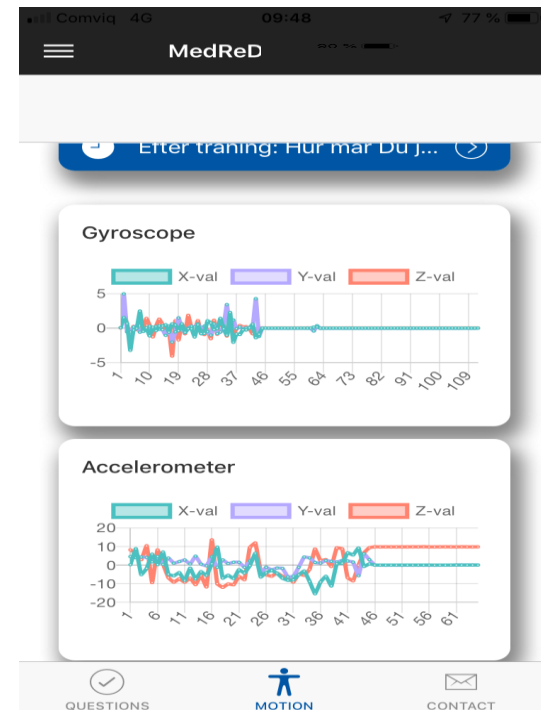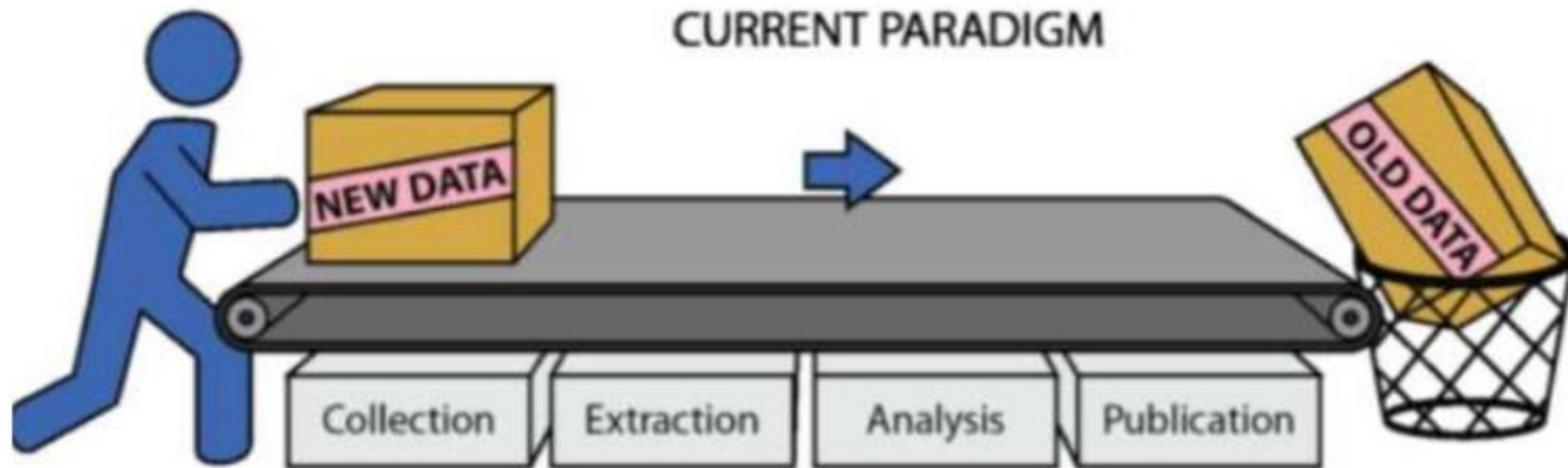
BAM by VizzDAT – "A LIMS IN POCKET"

o **BAM by VizzDAT is a mobile LIMS application, providing a cost-effective solution with superior ease of use, utilizing the most advanced modern data technology.**

o **Ideal for sample management in clinical studies.**

   o 100% customizable;

   o Easy and secure mobile application also available as desktop program. Compatible with all mobile devices (iOS, Android, Windows);

   o Enables storing information about biological samples and freezer positions, as well as a basic Electronic Medical Record (EMR) with sample donor and analysis data;

   o Seamless connectivity with various types of internal or external systems;

   o Facilitates instant delivery of e-reports to stakeholders;

   o Can be developed in multiple languages.

Connecting health and medical research with modern data technology

VizzDAT Platform Overview

Arbitrary External data sources

VizzDAT Applications

RaS /MedReD

BAM

Patients and Study Participants

Sample Logistics (e.g. research labs, biobanks)

Data Prep

Local data storage

VIZ - Interactive and integrated data visualisation tool

Mutation 91

Age

Data management

Data loss is real and significant, while data growth is staggering

@micheldumontier::BH17:2017-09-17

# Data loss is real and significant, while data growth is staggering

'Oops, that link was the laptop of my PhD student'

Limited Reproducibility of Scientific Medical Data

The Economist

OCTOBER 19TH–25TH 2013   Economist.com

Britain's angry white men
How to do a nuclear deal with Iran
Investment tips from Nobel economists
Junk bonds are back
The meaning of Sachin Tendulkar

HOW SCIENCE GOEs WRONG.

Too many of the findings that fill the academic ether are the result of shoddy experiments or poor analysis (see pages 21-24). A rule of thumb among biotechnology venture-capitalists is that half of published research cannot be replicated. Even that may be optimistic. Last year researchers at one biotech firm, Amgen, found they could reproduce just six of 53 "land-mark" studies in cancer research. Earlier, a group at Bayer, a drug company, managed to repeat just a quarter of 67 similarly important papers. A leading computer scientist frets that three-quarters of papers in his subfield are bunk. In 2000-10 roughly 80,000 patients took part in clinical trials based on research that was later retracted because of mistakes or improprieties.
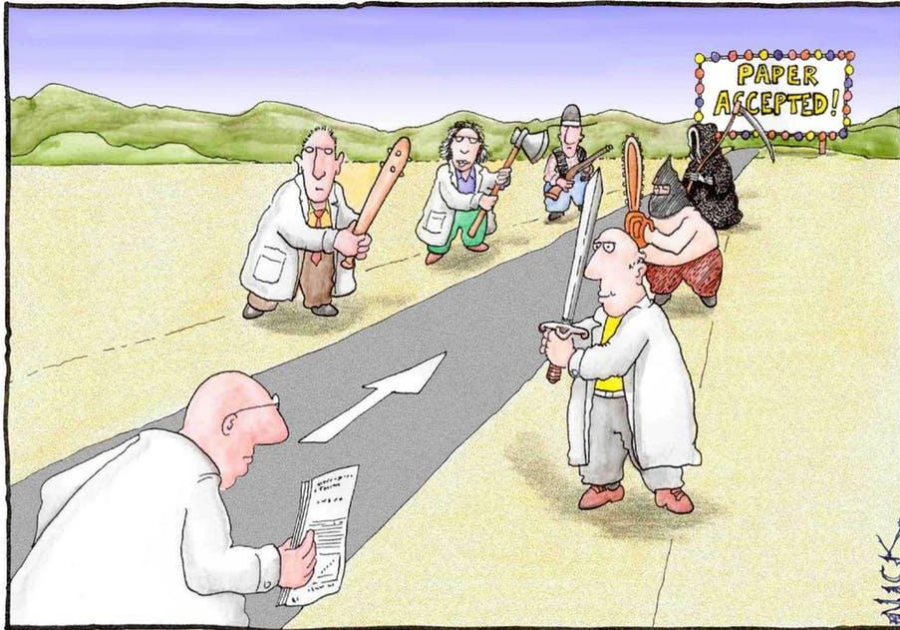
Reproducibility Crisis
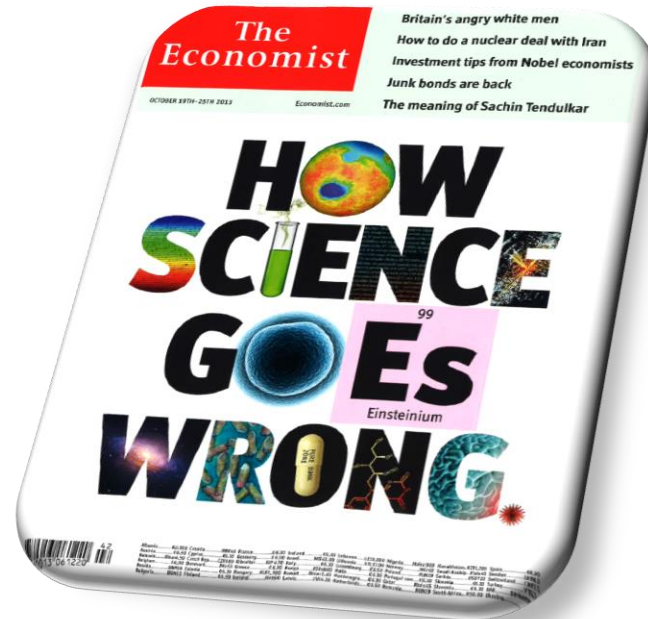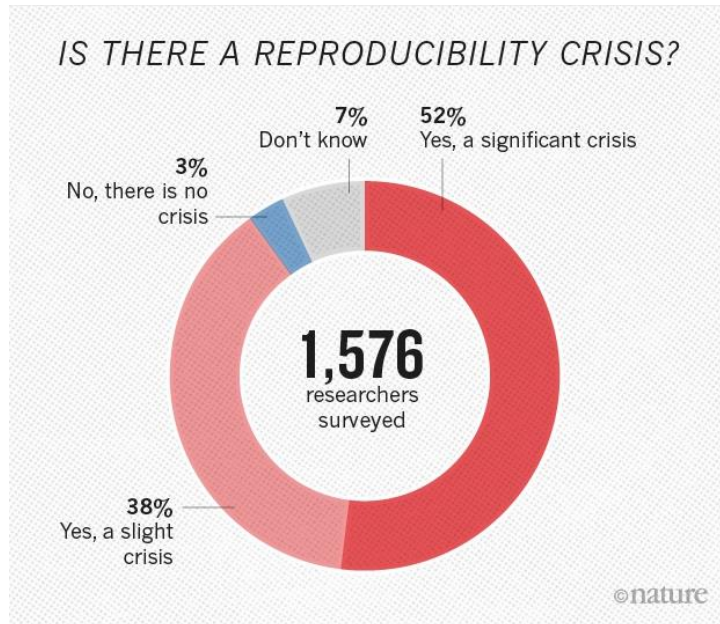
Reproducibility Crisis

TECHNOLOGY FEATURE · 20 AUGUST 2018

A toolkit for data transparency takes shape

A simple software toolset can help to ease the pain of reproducing computational analyses.

Moving up the academic ladder

Most scientists regarded the new streamlined peer-review process as "quite an improvement."

Reproducibility Crisis

One study found that only half of all reagents mentioned in over 200 recent articles from a range of journals and fields could be adequately identified, indicating a failure of researchers to comprehensively report the reagents they use and of editors and reviewers to require such reporting.

*Nicole A. Vasilevsky , Matthew H. Brush, Holly Paddock, Laura Ponting, Shreejoy J. Tripathy, Gregory M. LaRocca,
and Melissa A. Haendel (2013) _ On the reproducibility of science: unique identification of research resources in the biomedical literature,_ PeerJ, 1, e:148.*
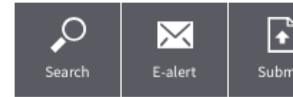
Reproducibility Crisis

nature
International journal of science

MENU ∨

Search   E-alert   Submit

TECHNOLOGY FEATURE · 20 AUGUST 2018

A to

A simpl

nature
International journal of science

MENU ∨

Search   E-alert   Subm

NEWS · 27 AUGUST 2018

High-profile journals put to reproducibility test

Researchers replicated 62% of social-behaviour findings published in Science and Nature — a result matched almost exactly by a prediction market.

The problem we don´t talk about!!

Ownership of data

The politics of data ownership and the lack of confidence in the complex synchronization that this requires has often stalled projects before they have even started.
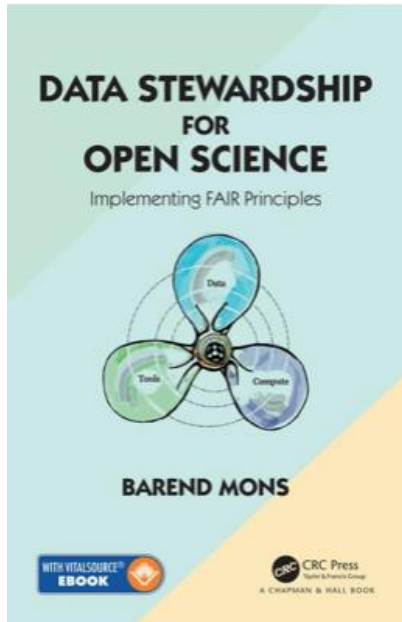
The solution, FAIR for e-Health?

F indable   A ccessible   I nteroperable   R eusable

A set of principles that apply to *all* digital resources

software, images, data, repositories, web services, scholarly publications

Data Stewardship for Open Science: Implementing FAIR Principles

1st Edition

Barend Mons

Hardback £80.00	Book + eBook £37.59	eBook £37.59	eBook Rental from £23.50

Chapman and Hall/CRC

Published March 5, 2018

Reference - 226 Pages - 19 B/W Illustrations

ISBN 9780815348184 - CAT# K345532

**For Instructors**   Request Inspection Copy

For Librarians   Available on Taylor & Francis eBooks >>

FAIR Data Action Plan

Interim recommendations and actions from the European Commission
Expert Group on FAIR data

June 2018

Sandra Collins, National Library of Ireland
Françoise Genova, Observatoire Astronomique de Strasbourg
Natalie Harrower, Digital Repository of Ireland
Simon Hodson, CODATA, Chair of the Group
Sarah Jones, Digital Curation Centre, Rapporteur
Leif Laaksonen, CSC-IT Center for Science
Daniel Mietchen, Data Science Institute, University of Virginia
Rūta Petrauskaité, Vytautas Magnus University
Peter Wittenburg, Max Planck Computing and Data Facility
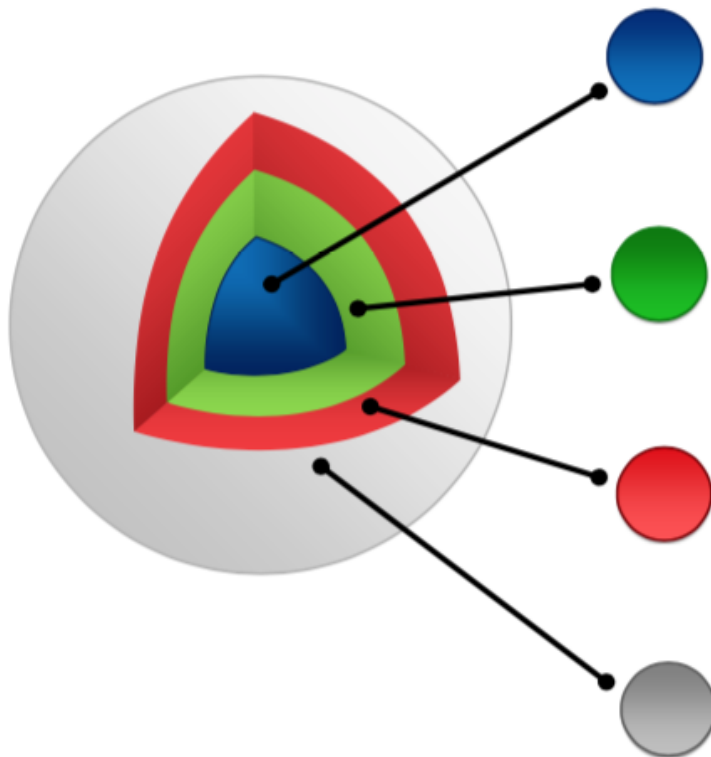
DATA
The core bits
At its most basic level, data is a bitstream or binary sequence. For data to have meaning and to be FAIR, it needs to be represented in standard formats and be accompanied by Persistent Identifiers (PIDs), metadata and code. These layers of meaning enrich the data and enable reuse.

IDENTIFIERS
Persistent and unique (PIDs)
Data should be assigned a unique and persistent identifier such as a DOI or URN. This enables stable links to the object and supports citation and reuse to be tracked. Identifiers should also be applied to other related concepts such as the data authors (ORCIDs), projects (RAIDs), funders and associated research resources (RRIDs).

STANDARDS & CODE
Open, documented formats
Data should be represented in common and ideally open file formats. This enables others to reuse the data as the format is in widespread use and software is available to read the files. Open and well-documented formats are easier to preserve. Data also need to be accompanied by the code use to process and analyse the data.

METADATA
Contextual documentation
In order for data to be assessable and reusable, it should be accompanied by sufficient metadata and documentation. Basic metadata will enable data discovery, but much richer information and provenance is required to understand how, why, when and by whom the data were created. To enable the broadest reuse, data should be accompanied by a 'plurality of relevant attributes' and a clear and accessible data usage license.

FAIR Data

Is FAIR enough to address e-Health research challenges?

FAIR Principles

Generic rules to make data:

Findable
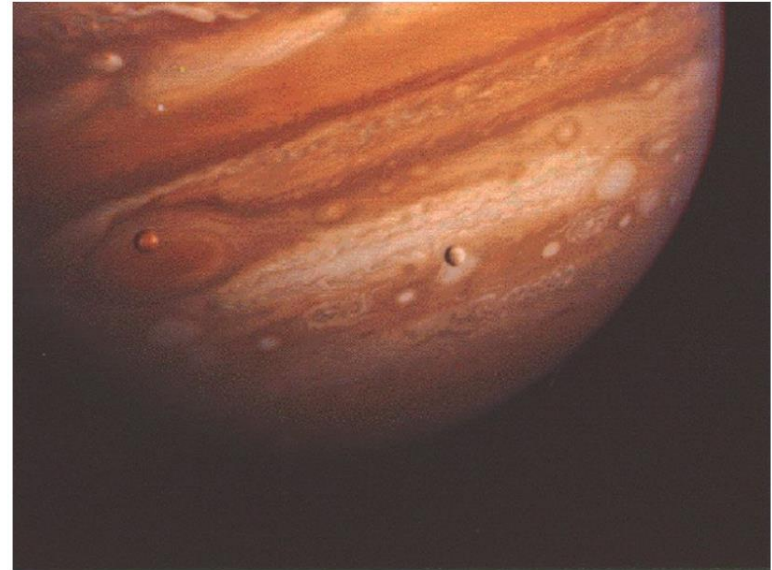
Accessible

Interoperable

Reusable

F

New Findings from Old Data

Recalibrated and reanalyzed data from the Voyager flybys of Jupiter 40 years ago, presented in a series of papers in *JGR: Space Physics*, show the value of archival data.
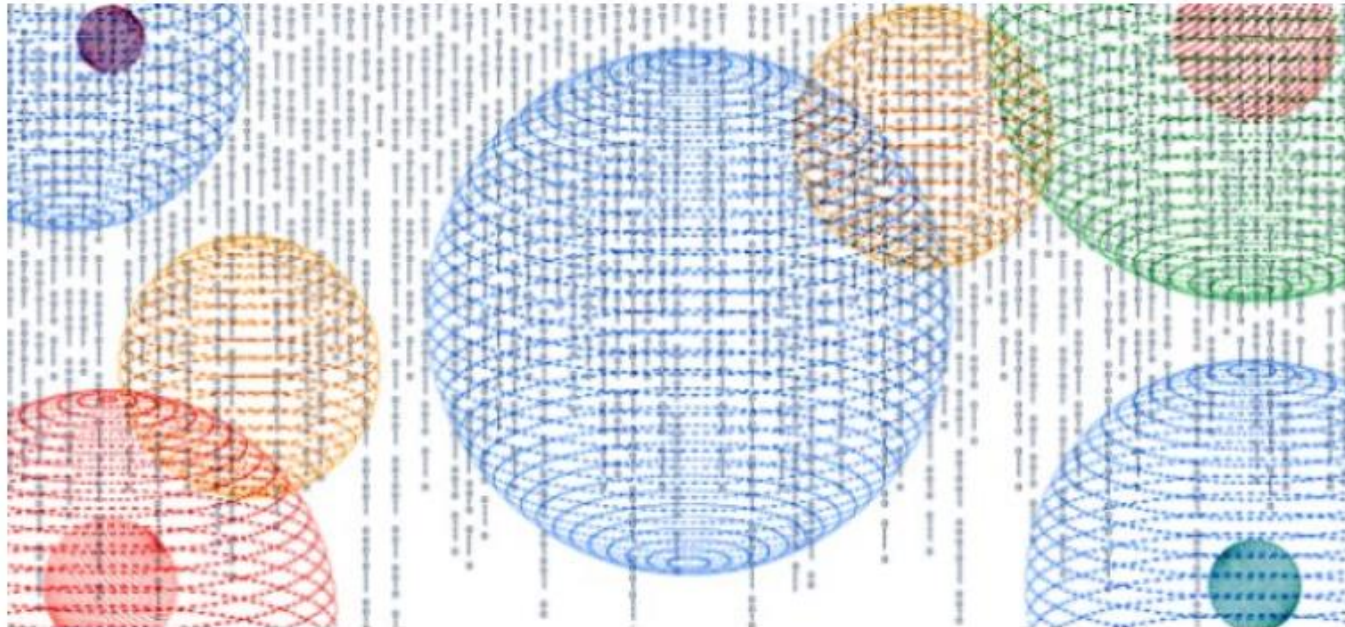

One of more than 33,000 pictures of Jupiter and its five major satellites taken by two Voyager spacecraft in 1979. Credit: NASA

By Mike Liemohn on 29 August 2017

NIH Launches Biomedical Data Ecosystem on Google Cloud

by Bruce Brown | August 17, 2018 | Enabling Tech, Health, Medical | 0 comments
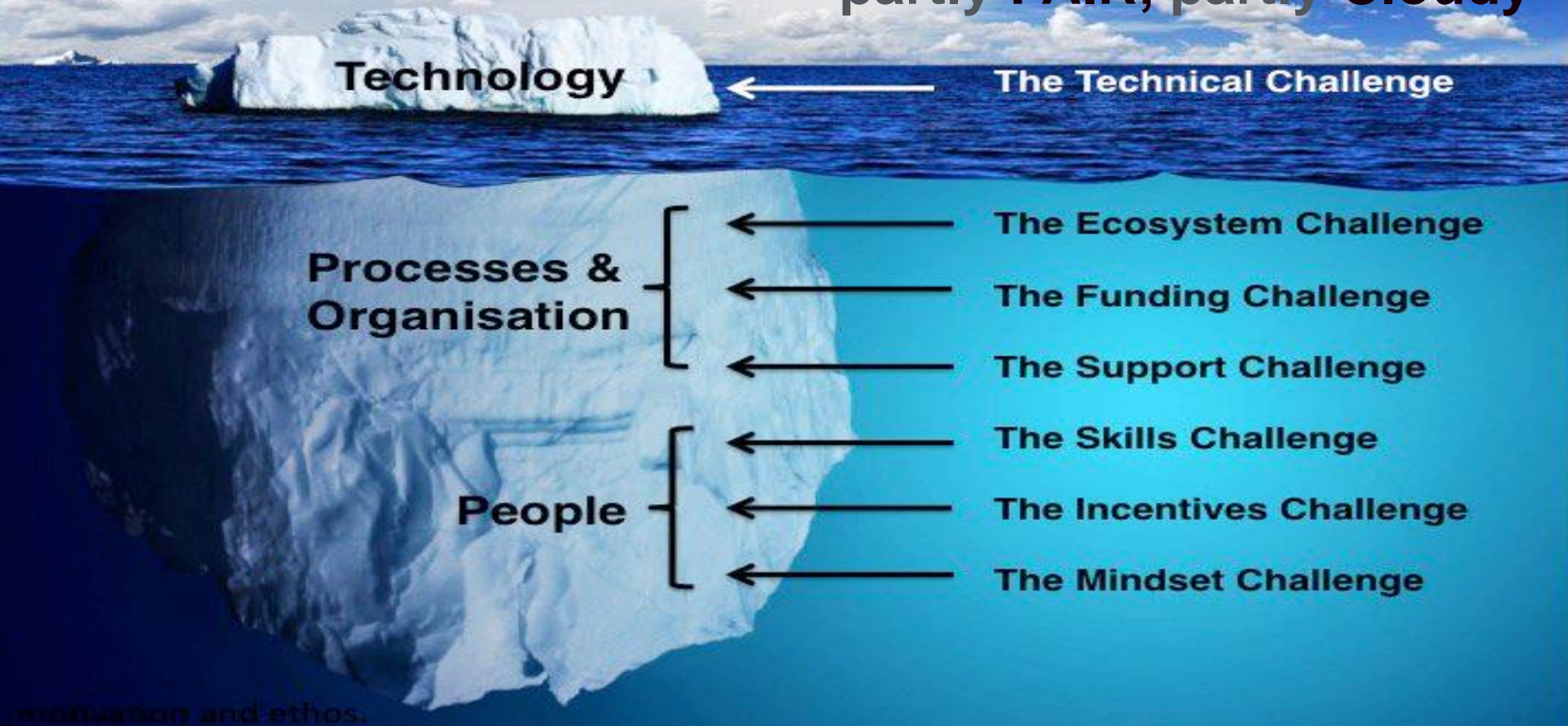
EOSC: **Challenges and Observations**

- The majority of the challenges are **social** rather than **technical**

- Not just t**he size of data,** but in particular **complex data** and **analytics**

- Shortage of **data experts** globally and in the European U

- **Archaic system of rewards** and **funding**

- '**Valley of death**' between **(e-)** **main specialists.**

- **Short funding cy** **ctures** are **not fit for purpose**

- **Fragmentation** be **causes** **repetitive** and **isolated** solutions

- Distributed data sets increasingly **do not move** (**size & privacy** reasons)

- Centralised HPC is **insufficient** to support **distributed meta-analysis and learning**.

- However, the **major components** for a **first generation EOSC** are largely 'there'

- But '**lost in fragmentation**' and spread over 28 Member States.

80% social/ 20% technical

The Open Data Iceberg

partly FAIR, partly Cloudy

Technology — The Technical Challenge

Processes & Organisation
- The Ecosystem Challenge
- The Funding Challenge
- The Support Challenge

People
- The Skills Challenge
- The Incentives Challenge
- The Mindset Challenge

motivation and ethos.

Developed from: Deetjen, U., E. T. Meyer and R. Schroeder (2015).

37

FAIR Principles

Findable

(F1) (meta)data are assigned a globally unique and eternally persistent identifier,

(F2) data are described with rich metadata,

(F3) (meta)data are registered or indexed in a searchable resource,

(F4) metadata specify the data identifier;

Annotation

80% of the rare diseases have genetic causes!

Annotation

- 27 % of all mutation in databases are wrongly annotated for Marfan Syndrom

Difficulties in diagnosing Marfan syndrome using current FBN1 databases.

Groth KA[1], Gaustadnes M[2], Thorsen K[2], Østergaard JR[3], Jensen UB[4], Gravholt CH[5], Andersen NH[6].

⊕ **Author information**

Abstract
PURPOSE: The diagnostic criteria of Marfan syndrome (MFS) highlight the importance of a FBN1 mutation test in diagnosing MFS. As genetic sequencing becomes better, cheaper, and more accessible, the expected increase in the number of genetic tests will become evident, resulting in numerous genetic variants that need to be evaluated for disease-causing effects based on database information. The aim of this study was to evaluate genetic variants in four databases and review the relevant literature.

METHODS: We assessed background data on 23 common variants registered in ESP6500 and classified as causing MFS in the Human Gene Mutation Database (HGMD). We evaluated data in four variant databases (HGMD, UMD-FBN1, ClinVar, and UniProt) according to the diagnostic criteria for MFS and compared the results with the classification of each variant in the four databases.

RESULTS: None of the 23 variants was clearly associated with MFS, even though all classifications in the databases stated otherwise.

CONCLUSION: A genetic diagnosis of MFS cannot reliably be based on current variant databases because they contain incorrectly interpreted conclusions on variants. Variants must be evaluated by time-consuming review of the background material in the databases and by combining these data with expert knowledge on MFS. This is a major problem because we expect even more genetic test results in the near future as a result of the reduced cost and process time for next-generation sequencing.Genet Med advance online publication 26 March 2015Genetics in Medicine (2015); doi:10.1038/gim.2015.32.

What was missing?

- The FAIR-Health principles for e-Health

- *Quality*
- *Incentive*
- *Provenance*

Recent studies in the field of biomedicine show that findings from an alarming percentage of scientific papers in even the top journals cannot be reliably reproduced by other researchers.
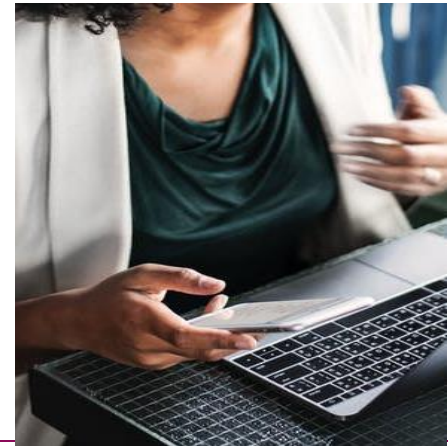
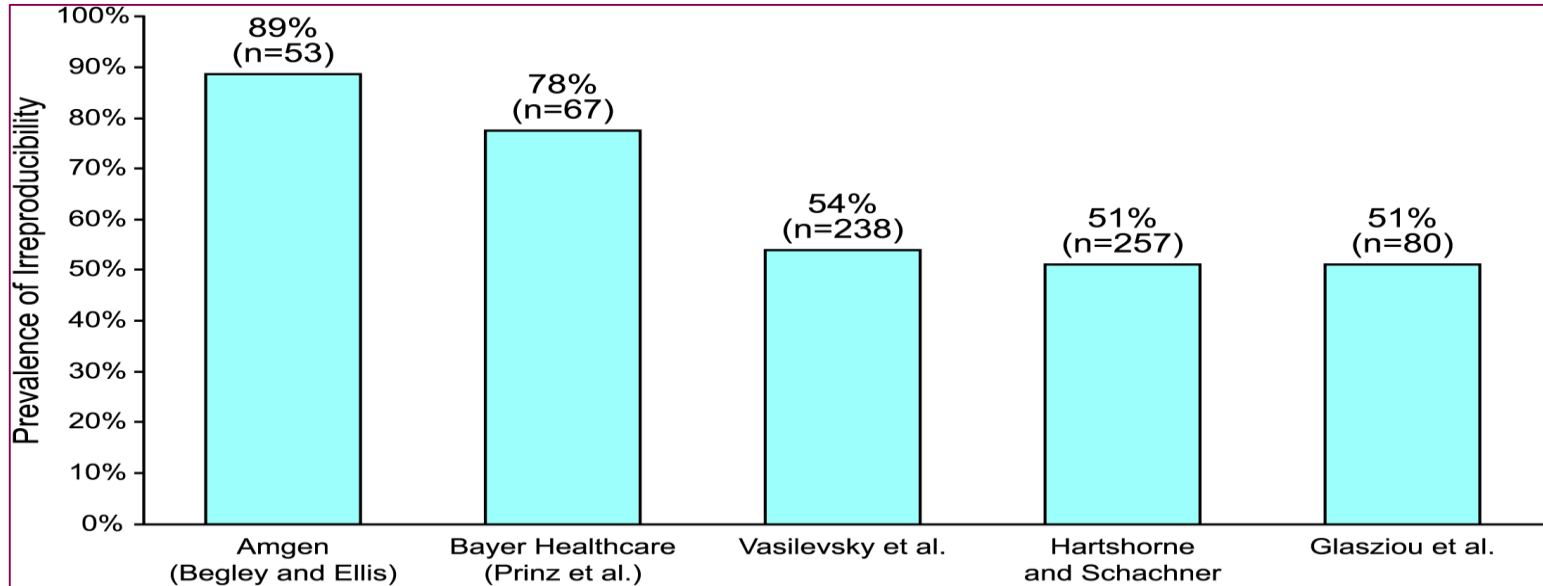# Some Medical Research Challenges Are Obvious – Prevalence Of Irreproducibility

Fig 1. Studies reporting the prevalence of irreproducibility. Source: Begley and Ellis [6], Prinz et al. [7], Vasilevsky [8], Hartshorne and Schachner [5], and Glasziou et al. [9].

One of the fundamental principles of science is *reproducibility* **– the idea that a discovery is valid only if any scientist in any lab can conduct the same experiment under the same conditions and obtain the same results. Without reproducibility, we could not distinguish scientific fact from error or chance, and scientific "laws" would vary around this planet.**
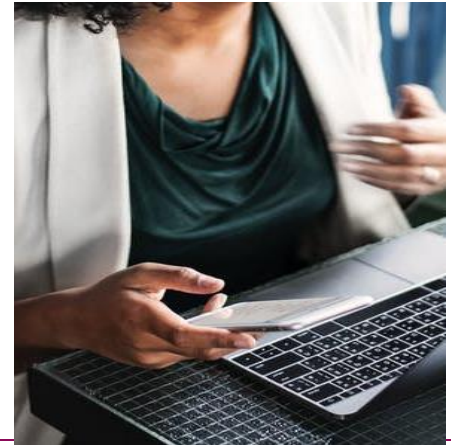
FAIR-Health to Help Medical Research and e-Health

We propose an extension of the FAIR Principles to include additional components:

- **Quality** aspects related to reproducibility and meaningful reuse of the data.
- **Provenance** information describing all steps.
- **Incentives** to stimulate effective enrichment of data sets and biological material collections and its reuse on all levels.
- **Privacy** respecting approaches for working with the biological material and data.

Data loss is real and significant, while data growth is staggering
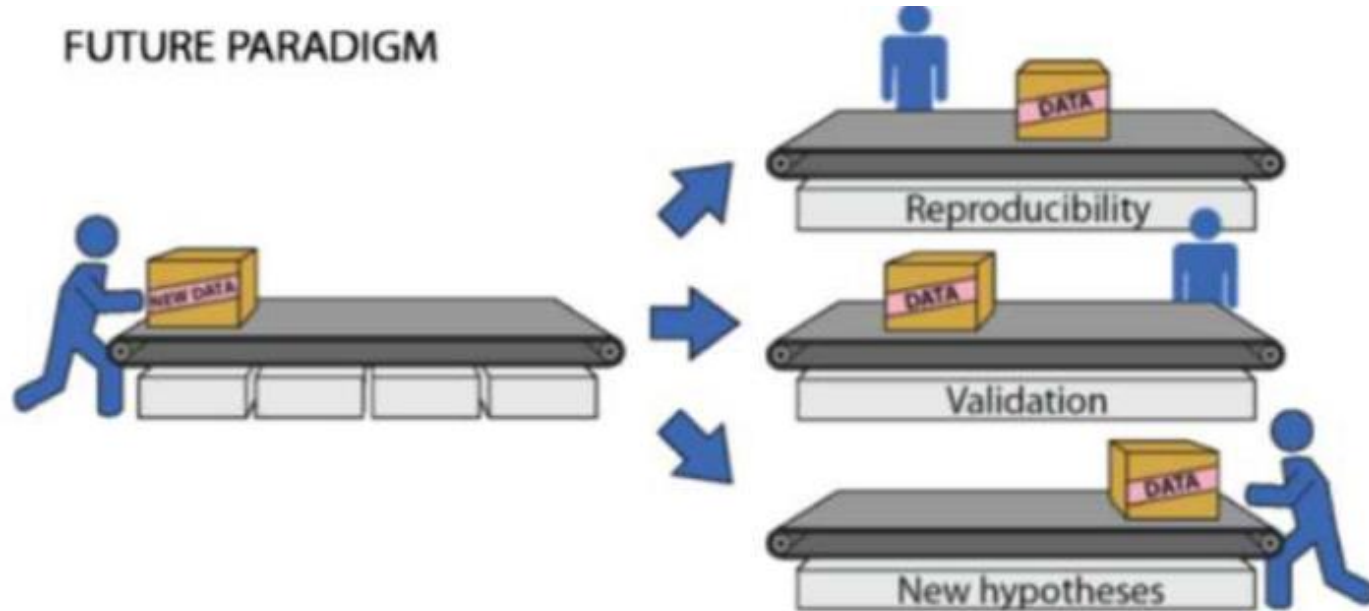
@micheldumontier::BH17:2017-09-17

Quality and Reproducibility

- Developing new CEN standards in Europe for biological material (SPIDIA4P)

- Developing ISO standards for biological material (ISO TC 276 - WG2)

- Developing provenance information standards in TC276 (WG5)

More to Come (SPIDIA4P) ……

- Venous whole blood — isolated circulating tumour cells, (CTCs) and circulating organ cells, (COCs), isolated DNA, RNA, proteins

- Venous whole blood – Isolated exosomes isolated nucleic acids

- Urine and other body fluids – isolated cfDNA

- Saliva – isolated human DNA

- Saliva and stool – isolated microbiome DNA

- Frozen Tissue – isolated DNA

- Fine Needle Aspirates (FNAs) – isolated DNA, RNA, proteins

- Metabolomics of body fluids: International ISO Standard: ISO/TC 212

- FFPE Tissue – in situ stainings including immunohistochemistry (IHC): ISO/TC 212

Provenance information

The Provenance of a piece of data refers to knowledge about is origin.

A major challenge in data-driven biomedical research lies in the collection and representation of data provenance information to ensure that findings are reproducibile.
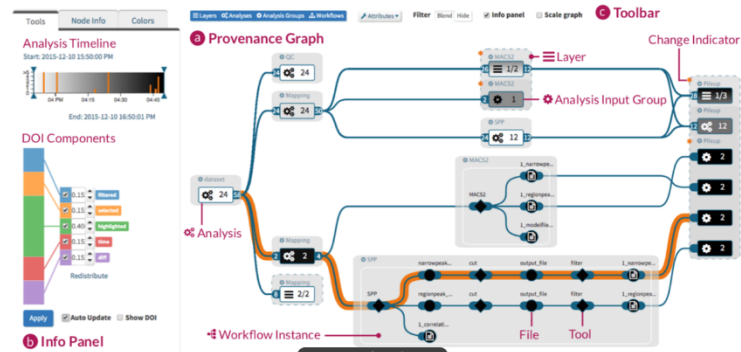

AVOCADO: Visualization of Workflow-Derived Data Provenance for Reproducible Biomedical Research

H. Stitz[1], S. Luger[1], M. Streit[1*], and N. Gehlenborg[2*]

[1]Johannes Kepler University Linz, Austria
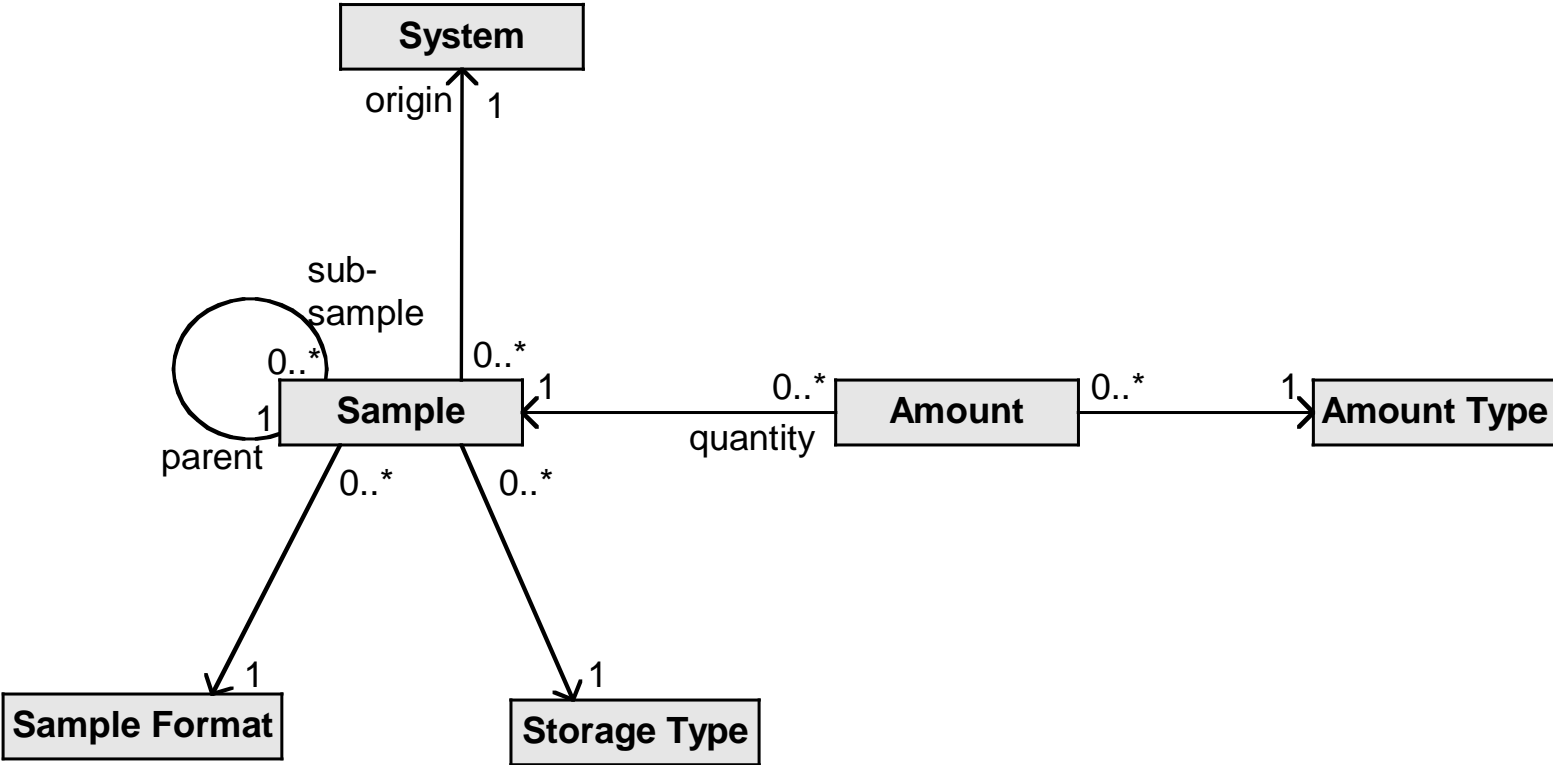[2]Harvard Medical School, United States of America
*Equal contribution

Provenance information

- Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness.

- This concept is used in archival science, archeology and paleontology, computer science, bioinformatics, business processes, and other domains.

Provenance information

- Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness.

- This concept is used in archival science, archeology and paleontology, computer science, bioinformatics, business processes, and other domains.

- Complete provenance information of any biological material and data is important in order to interpret the data or to enrich an existing biological material and data set consistently.

- This provenance information must include a link to the source biological material and—if possible—a link to the information on the very research participant who donated the material.
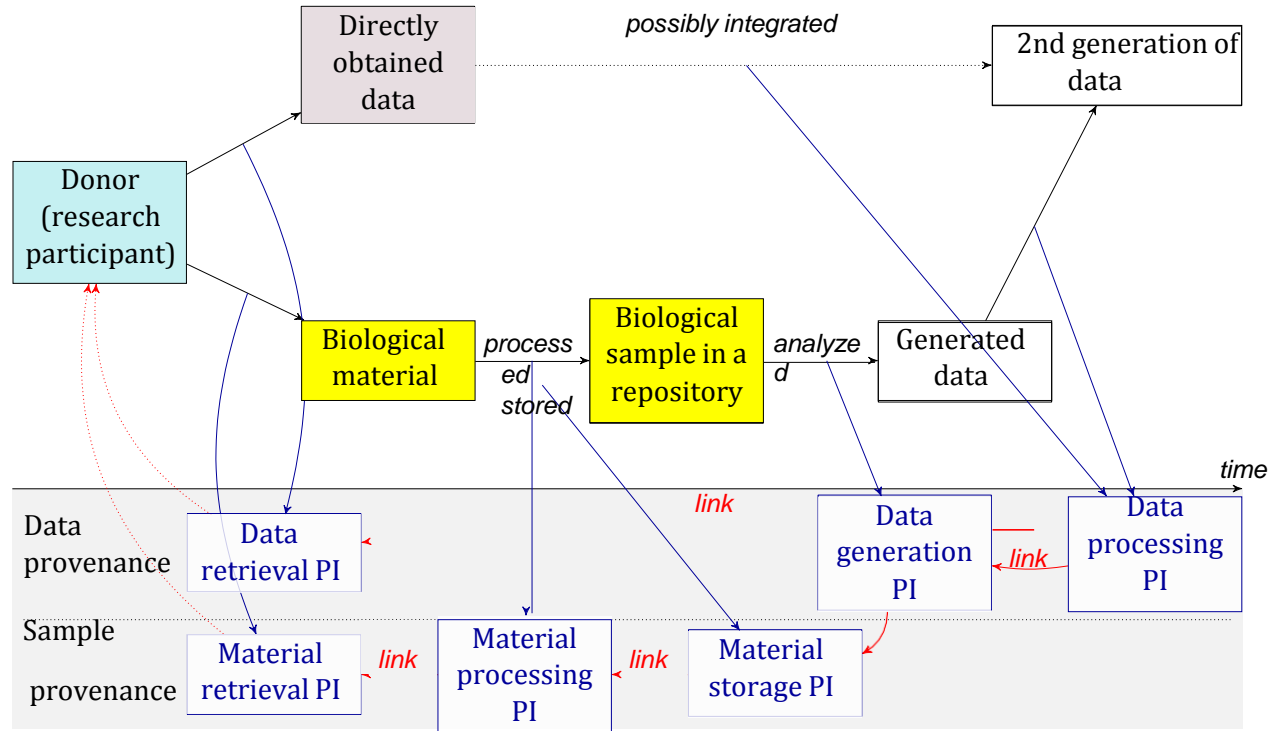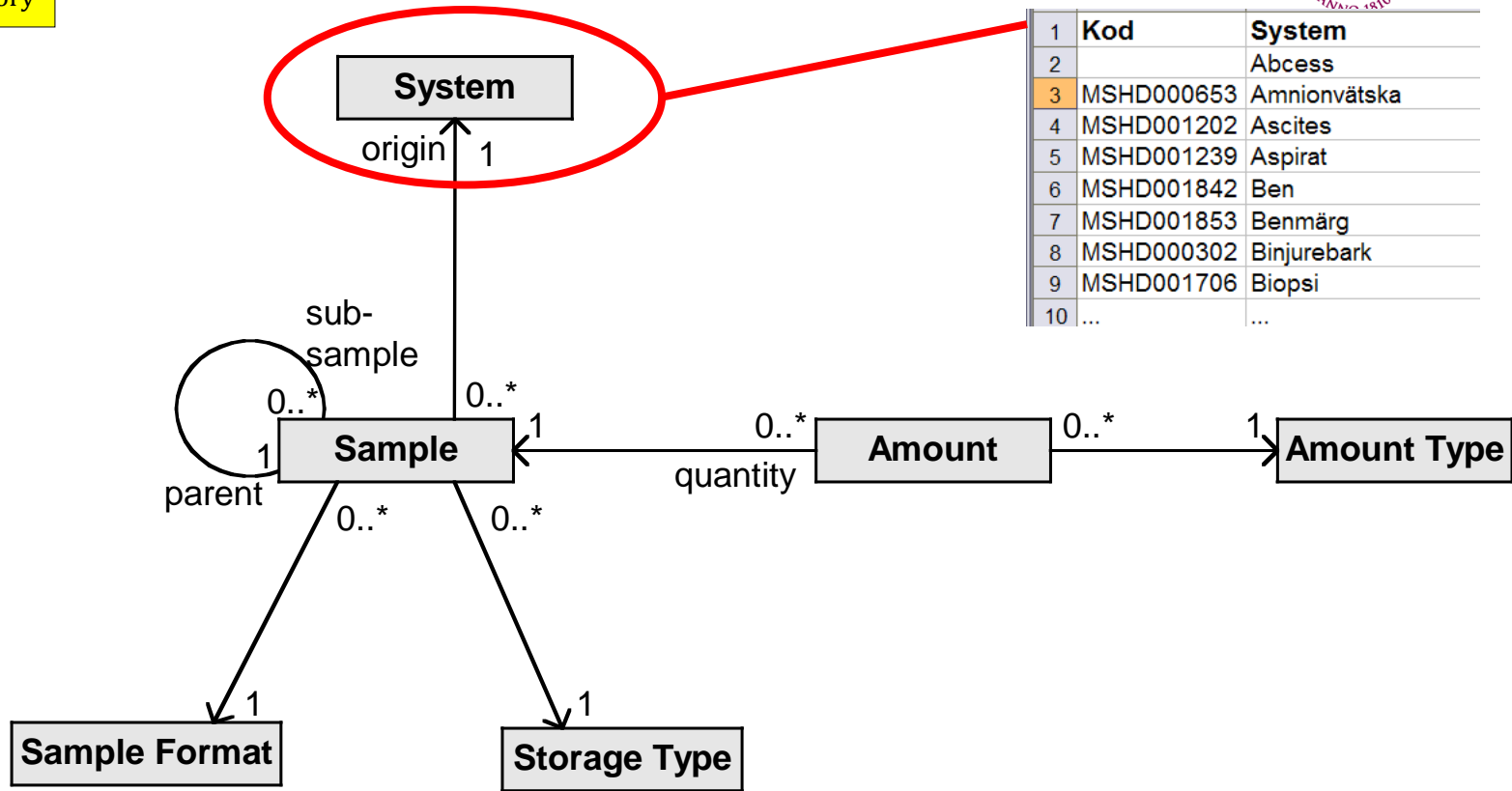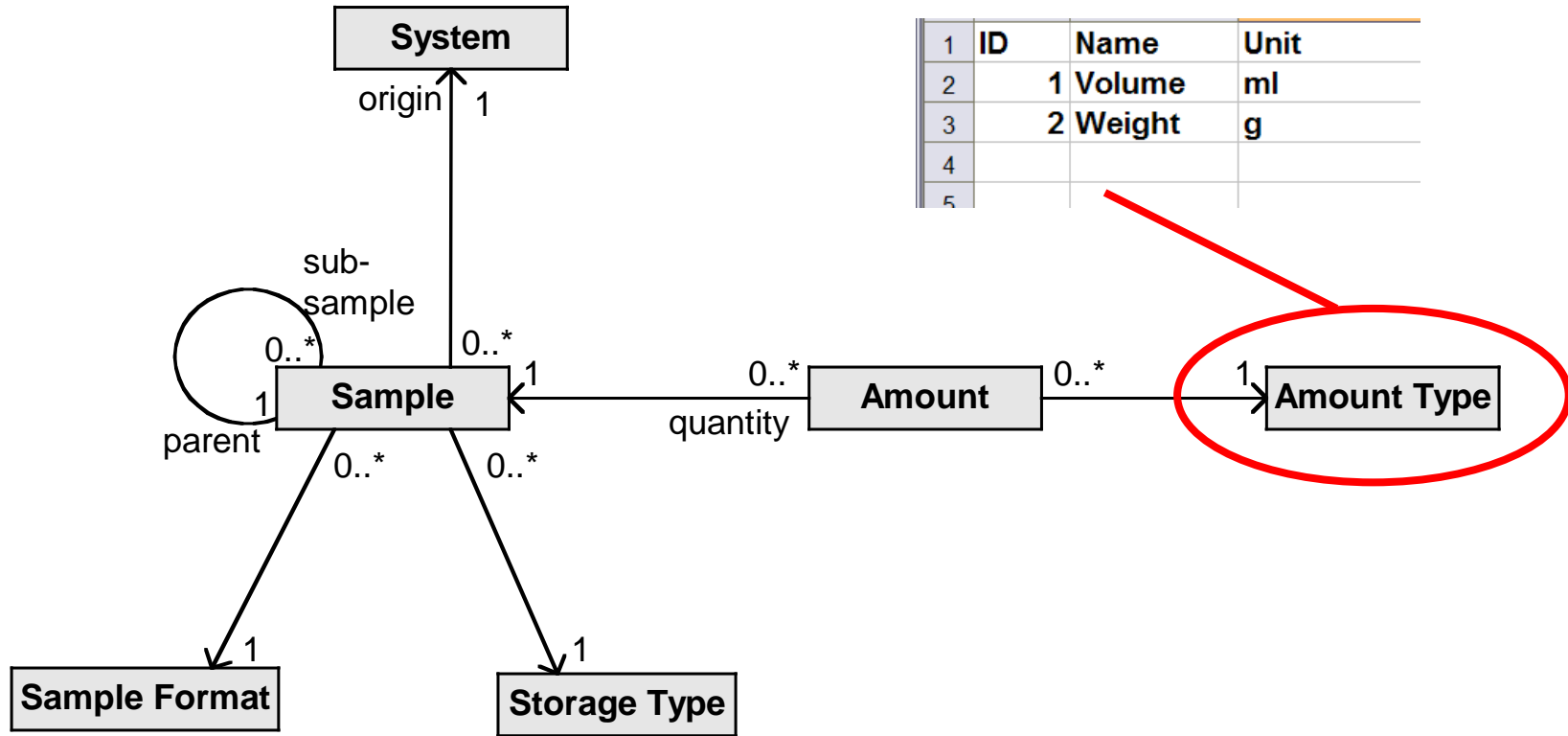
A coverage of provenance information for biological material and associated data. Provenance information is abbreviated as PI in the figure.
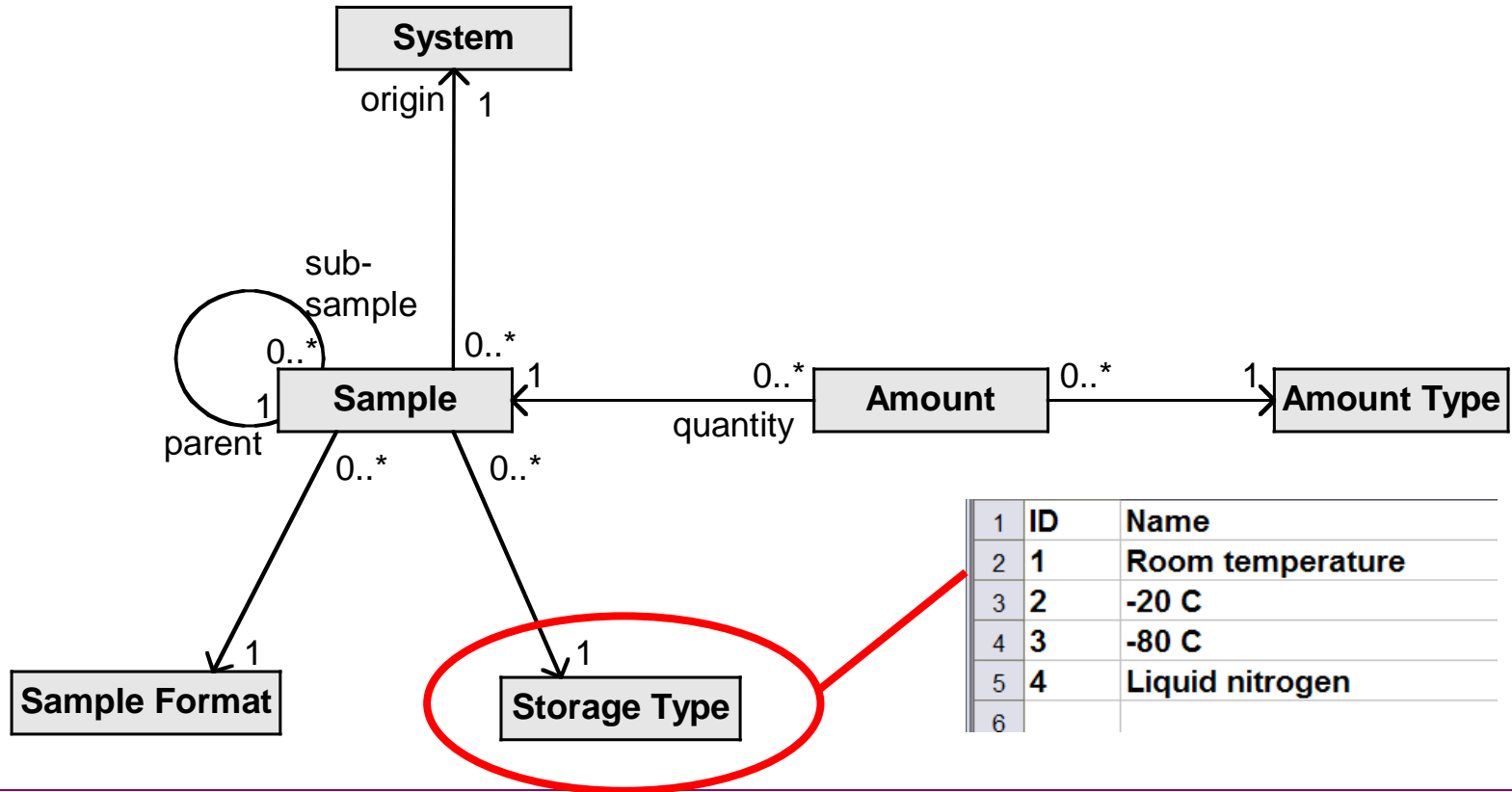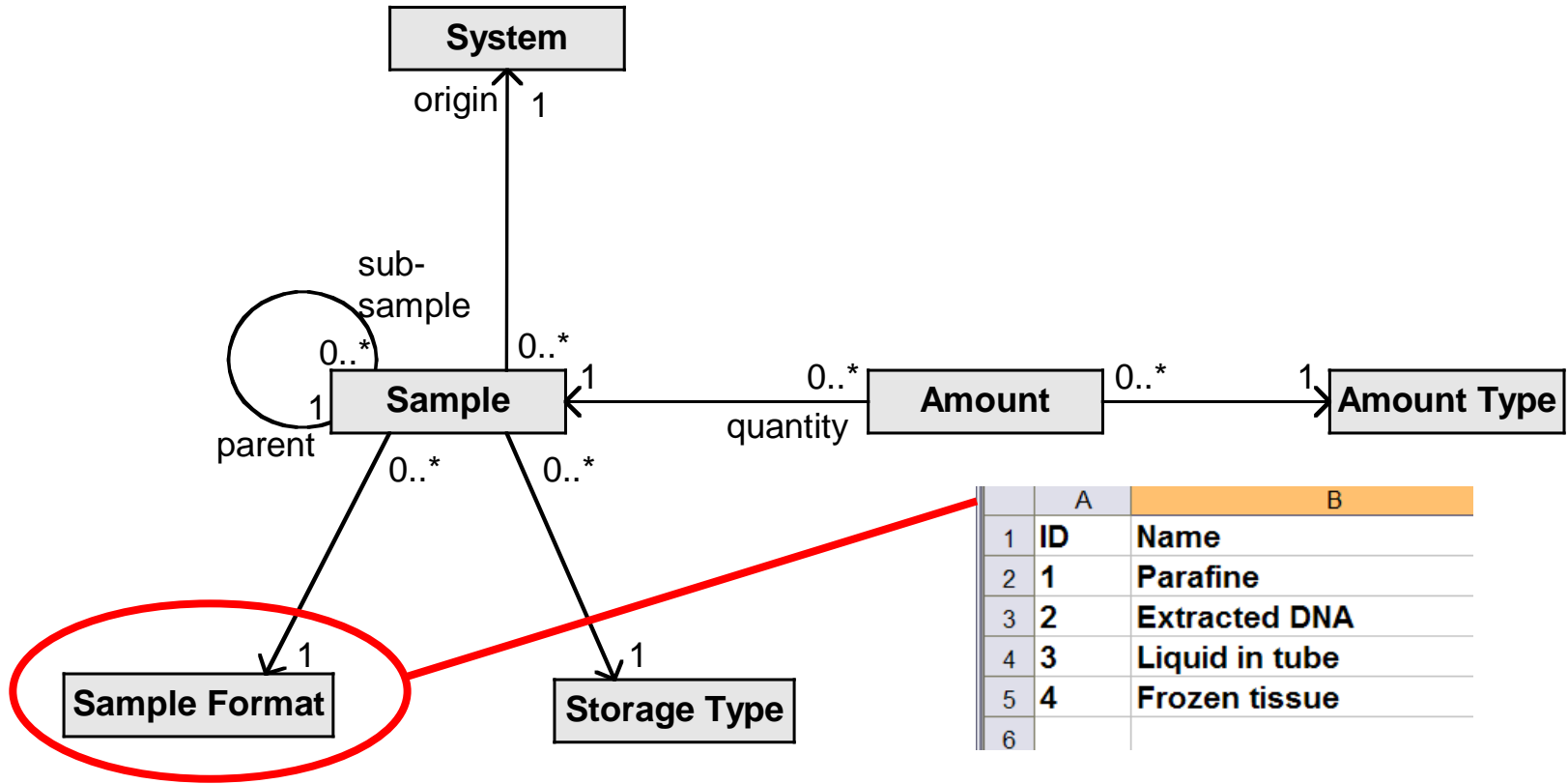
Physical entities
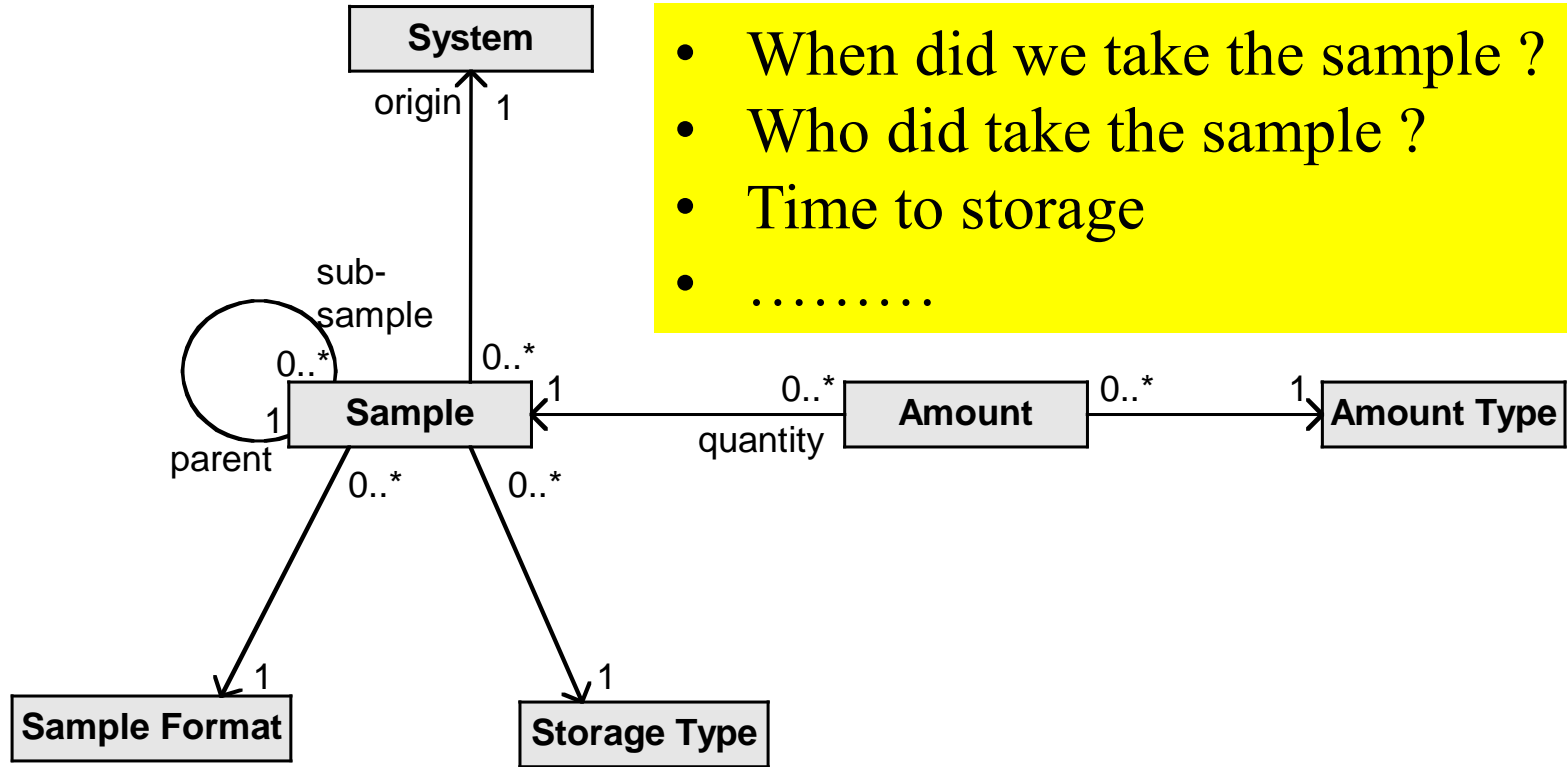
Biological sample in a repository

System

origin 1

	ID	Name	Unit
1			
2	1	Volume	ml
3	2	Weight	g
4			
5			

sub-sample

0..*

0..*

1 Sample

0..* Amount 0..* 1 Amount Type

quantity

1 parent

0..* 0..*

1 Sample Format

1 Storage Type

Karolinska Institutet

System

origin   1

sub-sample

0..*

0..*

1

Sample   1   0..*   **Amount**   0..*   1   **Amount Type**

1

parent   quantity

0..*   0..*

1   1

Sample Format   **Storage Type**

	ID	Name
1	ID	Name
2	1	Room temperature
3	2	-20 C
4	3	-80 C
5	4	Liquid nitrogen
6		

Karolinska Institutet

System

origin   1

sub-sample

0..*   0..*
                     1
1   Sample   quantity   0..*   Amount   0..*   1   Amount Type

parent

0..*   0..*

1   1
Sample Format   Storage Type

	A	B
1	ID	Name
2	1	Parafine
3	2	Extracted DNA
4	3	Liquid in tube
5	4	Frozen tissue
6		

Biological sample in a repository

System

origin 1

sub-sample

0..*

parent 1

Sample

0..*

0..*

0..*

quantity

0..* **Amount** 0..* 1 **Amount Type**

1

Sample Format

1

Storage Type

- When did we take the sample ?
- Who did take the sample ?
- Time to storage
- ………

Desiderata for Terminologies

- **Make sure the semantics are universally understood, separate from linguistics**

- **Make sure that, as our understanding changes, original meaning is not forgotten**

- **Provide a bridge between what we record and how we reason**

Incentives

- In contrast to many other scientific fields, medical data can only be made available to researchers because of voluntary contribution of citizen, in particular research participants (donors and patients).

- A positive incentive scheme must be developed and adopted in wide research communities, which will maximize biological material and data sharing and achieve actual *reuse*.

- incentive principles should also be applied to software tools and their sustainability, which is fundamental for any data-driven medical research.

Data Integration Challenge
Ownership of data

The politics of data ownership and the lack of confidence in the complex synchronization that this requires has often stalled projects before they have even started.

"Do You Expect Me to Just Give Away My Data?"

The Editor-in-Chief of *JGR: Oceans* explains why the new AGU data policy is important for the rigor and long-term security of scientific research.
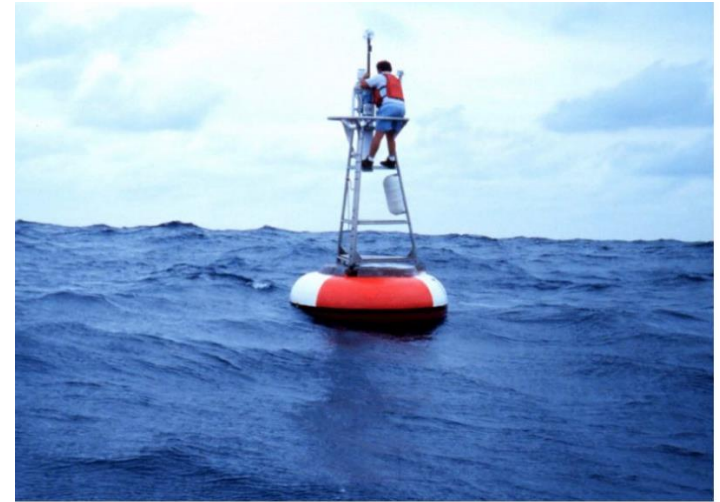
We appreciate that many researche Skärmavbild gths to collect their data but making data sets fully accessible to others via public repositories is important for the future of science. Credit: NOAA

FAIR- Health include Incentives

Privacy Protection

For human data used in biomedical research, there are three naturally competing interests:

- protection of privacy of individuals contributing their personal potentially privacy sensitive data

- reuse of data to maximize return on investment into research and society

- complex ownership situation and economic interests

These needs have been recognized by various medical communities, as witnessed by the efforts toward clinical trials data sharing

Finding human data
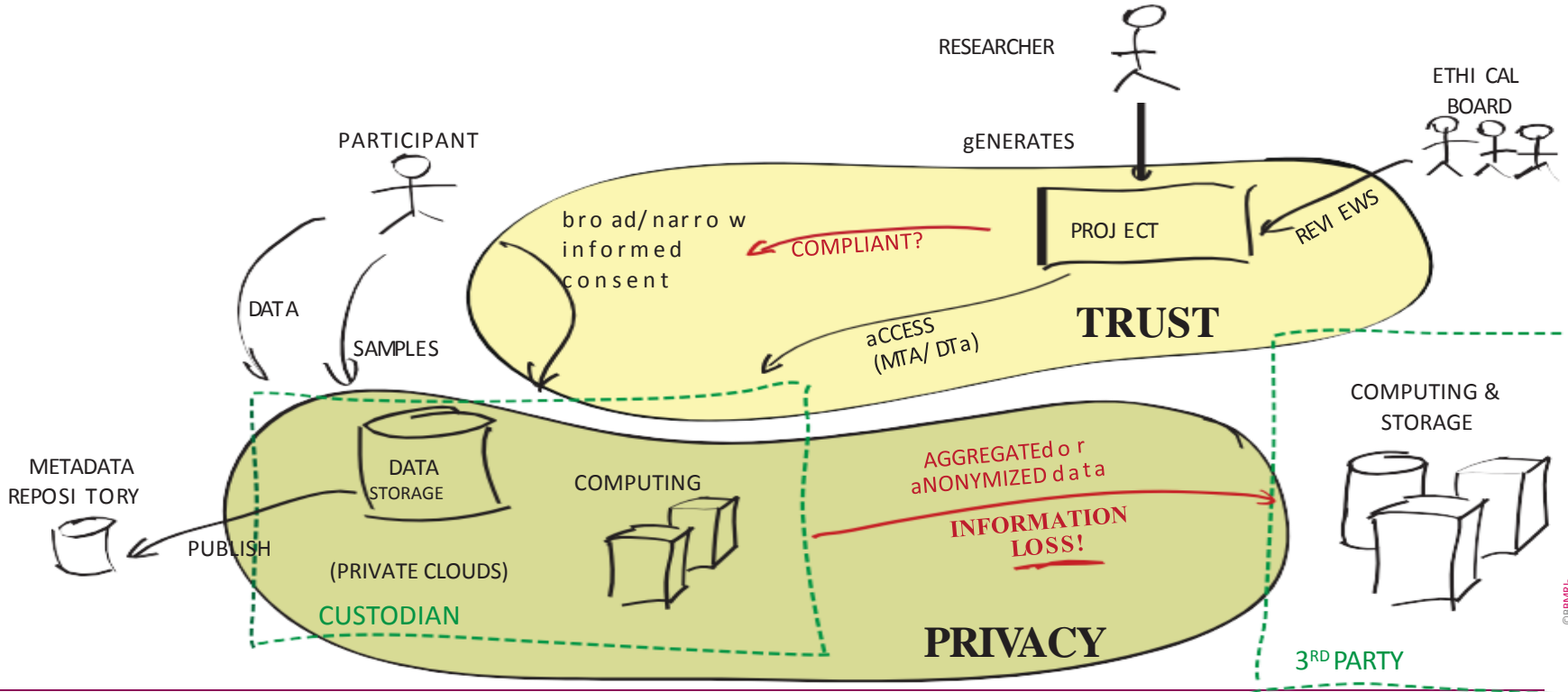
General Data Protection Regulation - 25 May 2018

Nature editorial

Concern: Health research is not specifically addressed


(a)   Conditions for consent
(b)   Secondary use of data
(c)   Personal data versus anonymised data
(d)   Defining and dealing with genetic data
(e)   Data/sample transfer to 3rd countries and international organisations
(f)   EOSC

- **<u>Interpretation</u> and implementation of the EU data protection framework could differ considerably**

Blockchain aims to fumdamentally changing the way sensitive data are shared.
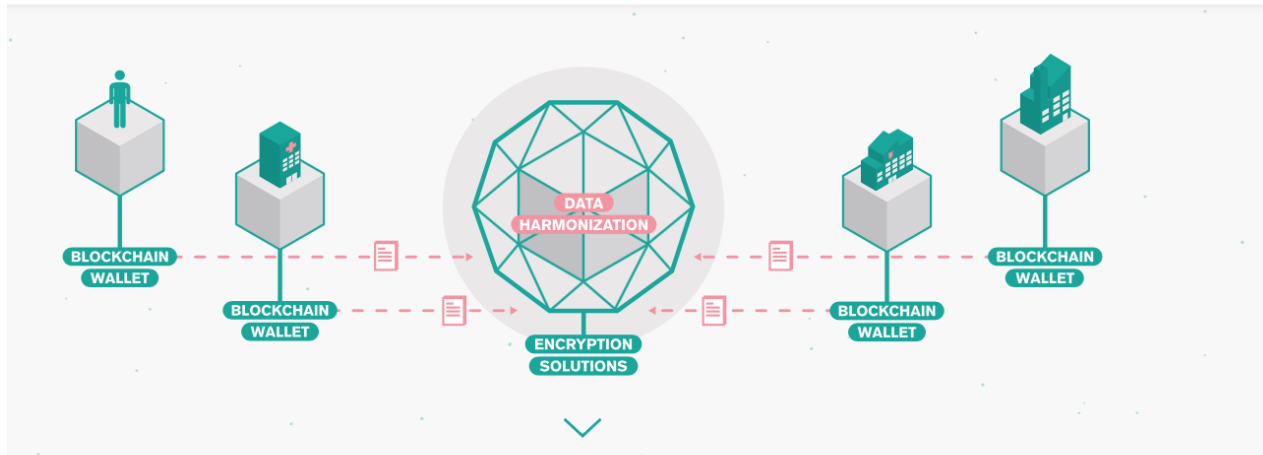
A NEW PARADIGM IN HEALTHCARE DATA PRIVACY AND SECURITY

MyHealthMyData (MHMD) is a Horizon 2020 Research and Innovation Action which aims at fundamentally changing the way sensitive data are shared. MHMD is poised to be **the first open biomedical information network centred on the connection between organisations and individuals**, encouraging hospitals to start making anonymised data available for open research, while prompting citizens to become the ultimate owners and controllers of their health data. MHMD is intended to become **a true information marketplace**, based on new mechanisms of trust and direct, value-based relationships between **EU citizens, hospitals, research centres and businesses**.

Blockchain aims to fumdamentally changing the way sensitive data are shared.

Will Blockchain Transform Healthcare?

Randy Bean Contributor
CIO Network Contributor Group ⓘ
Enterprise & Cloud

Consent management. In the current healthcare environment where every state has different privacy and consent regulations, blockchain could be used to record patient consent for purposes of data sharing.   Any party seeking to exchange medical data about a patient could check the blockchain for permission to do so.

Is FAIR enough to address medical research challenges?

Is FAIR enough to address medical research challenges?

NO !!

For medical research, all of these components called *FAIR-Health* are fundamental prerequisites for effective reuse of the biological material and data.

Enhancing *Reuse* of Data and Biological Material in Medical Research: From FAIR to FAIR-Health

Petr Holub[*1], Florian Kohlmayer[2], Fabian Prasser[2],
Michaela Th. Mayrhofer[1], Irene Schlünder[1,3], Gillian M. Martin[4],
Sara Casati[5], Lefteris Koumakis[6], Andrea Wutte[1], Łukasz
Kozera[7], Dominik Strapagiel[8], Gabriele Anton[9],
Gianluigi Zanetti[10], Osman Ugur Sezerman[11], Maimuna
Mendy[12], Dalibor Valík[13], Marialuisa Lavitrano[5], Georges
Dagher[14], Kurt Zatloukal[15], GertJan B. van Ommen[16], and
Jan-Eric Litton[1]

jan-eric.litton@ki.se