

Data integrity and data management

Data integrity – or: “Can you repeat it?”

Your experiment
from 2 years ago?



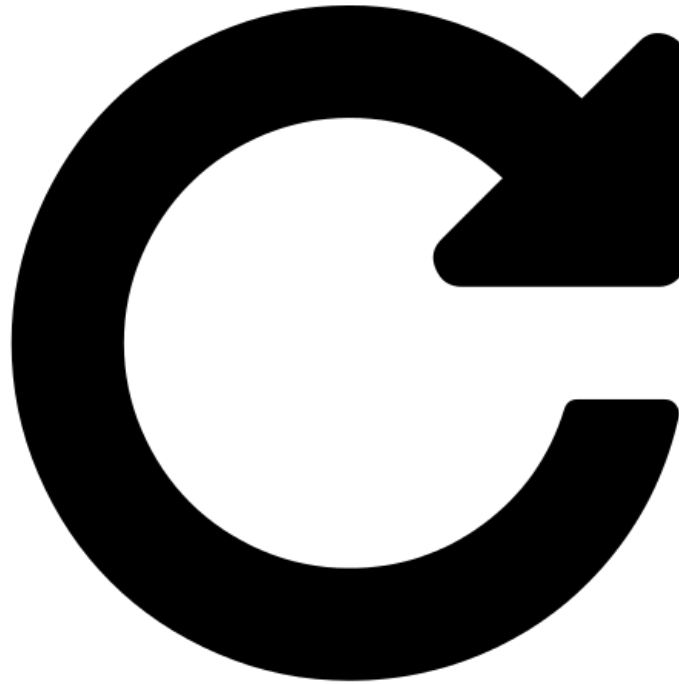
Your colleagues' experiment
from 2 years ago?

Your experiment in
2 years?

Data integrity – a definition

Data integrity

- degree to which data are **complete, consistent, accurate, trustworthy** and **reliable**
- these characteristics are **maintained** throughout the **data life cycle**



Good Documentation Practice

- methods for creating and maintaining data records, to ensure data integrity throughout the data life cycle

All **original records** and **documentation**, which are the result of the observations and activities in a study.

Raw data may include:

- photographs, videotapes, blots, chromatograms, computer readable media, dictated observations, ...
- data directly entered into a computer through an automatic instrument interface
- copies of original laboratory records and documentation that are complete and of good quality.

Processed result may also be recognized as raw data if original observations cannot be stored for technical reasons, e.g.:

- when data sets are recorded in a specific format that may or may not be readable at a later time point
- when exceptionally large volumes of data are generated that are technically difficult to store without pre-processing to reduce the storage volume

Increasing awareness of data integrity issues

GlaxoSmithKline Fires China R&D Boss for 'Misrepresented' Data

Dan Mangan | @_DanMangan
Tuesday, 11 Jun 2013 | 6:35 PM ET



Francois Lo Presti | AFP | Getty Images

A worker in vaccine production at drugmaker GlaxoSmithKline in Saint-Amand-les-Eaux, France.

- R&D boss fired
- Researcher resigned
- Paper retracted

“
”

***It takes 20 years to build a reputation
and five minutes to ruin it.
If you think about that, you'll do things
differently.***

Warren Edward Buffett, born 1930, American business
magnate, investor and philanthropist



The two principles

ALCOA plus

FAIR

Acronym	Attributable Legible Contemporaneous Original Accurate	Complete Consistent Enduring Available	Findable	Accessible
			Interoperable	Reusable
Focus	<ul style="list-style-type: none">• On the experiment• On describing attributes directly related to the data		<ul style="list-style-type: none">• Beyond the experiment, when making data broadly available• On the description and presentation of meta data	
Origin	FDA around 1968, WHO guidance		Lorentz Center workshop in Leiden in the Netherlands in 2015	
Reference	Section 9 Good Documentation Practice		Published as a <i>Nature</i> article	

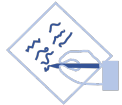
ALCOA and ALCOA plus

ALCOA



Attributable

Who acquired the data?



Legible

Can you read and understand the data?



Contemporaneous

Was it recorded as it happened?



Original

Are raw data records saved?



Accurate

Are all the details correct?

plus



Complete

Are all data included?



Consistent

Are all elements in chronological order?



Enduring

Are all recordings and notes permanent?



Available

Can the data be accessed any time?

Findable

Accessible

Interoperable

Reusable

Make scientific data FAIR

All disciplines should follow the geosciences and demand best practice for publishing and sharing data, argue Shelley Stall and colleagues.

Shelley Stall , Lynn Yarmey, Joel Cutcher-Gershenfeld, Brooks Hanson, Kerstin Le
Lesley Wyborn



Find

Inter



Oceanographer David Holland repairs a broken navigation module in Greenland. Credit: Lucas Jackson/Reuters



Findable

Accessible

Interoperable

Reusable

FAIR - Findable

Findable

- Global unique ID
- by humans and computers

Accessible

Interoperable

Reusable

FAIR - Accessible

Findable

- Global unique ID
- by humans and computers

Accessible

- The path to access it
- Authorisation and authentication

Interoperable

Reusable

FAIR – Interoperable

Findable

- Global unique ID
- by humans and computers

Accessible

- The path to access it
- Authorisation and authentication

Interoperable

- Usable on other systems

Reusable

FAIR - Reusable

Findable

- Global unique ID
- by humans and computers

Accessible

- The path to access it
- Authorisation and authentication

Interoperable

- Usable on other systems

Reusable

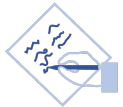
- Enough meta data

The principles

ALCOA



Attributable



Legible



Contemporaneous



Original



Accurate

plus



Complete



Consistent



Enduring



Available

FAIR

Findable

Accessible

Interoperable

Reusable

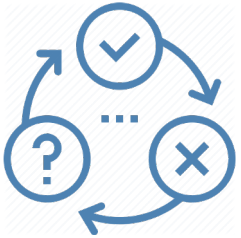
A few suggestions to increase data integrity



Establish unique study IDs



Use templates as often as possible



Establish a routine



Create reproducible protocols

<https://www.nature.com/articles/d41586-021-02428-3>



Align between lab members



Use a data management plan

Questions

Do you think you can apply ALCOA and FAIR?

Do you see any challenges?

What would be the biggest challenge?