Data management hits the workfloor

Opportunities, Obstacles and Challenges

Bela Mulder

AMOLF physics of functional complex matter

Why am I here?

My claim to datamanagement fame

- Member “Klankbordgroep Beleidskader Datamanagement NWO-instituten.” 2016-2017
- Member team that wrote the AMOLF Datamanagement plan 2017-2018
- Chairperson working group Replication Packages AMOLF 2018

Early 2015

Beleidskader Datamanagement NWO-Instituten

Introductie
NWO streeft optimale toegankelijkheid na van de resultaten van door haar gefinancierd onderzoek, *Open Science* (NWO Strategie 2015-2018). Het beleidskader datamanagement NWO-instituten is een uitwerking van de bredere NWO strategie op het gebied van Open Science, waaronder ook Open Access van wetenschappelijke publicaties en Wetenschappelijke Integriteit worden begrepen. Voor publicaties is reeds een open access-beleid ontwikkeld en geïmplementeerd, o.a. via de Regeling Subsidies. In de NWO-subsidieprogramma's wordt voor data bij indiening gevraagd om een dataparagraaf op te stellen en bij toekenning moet een datamanagementplan worden opgesteld.

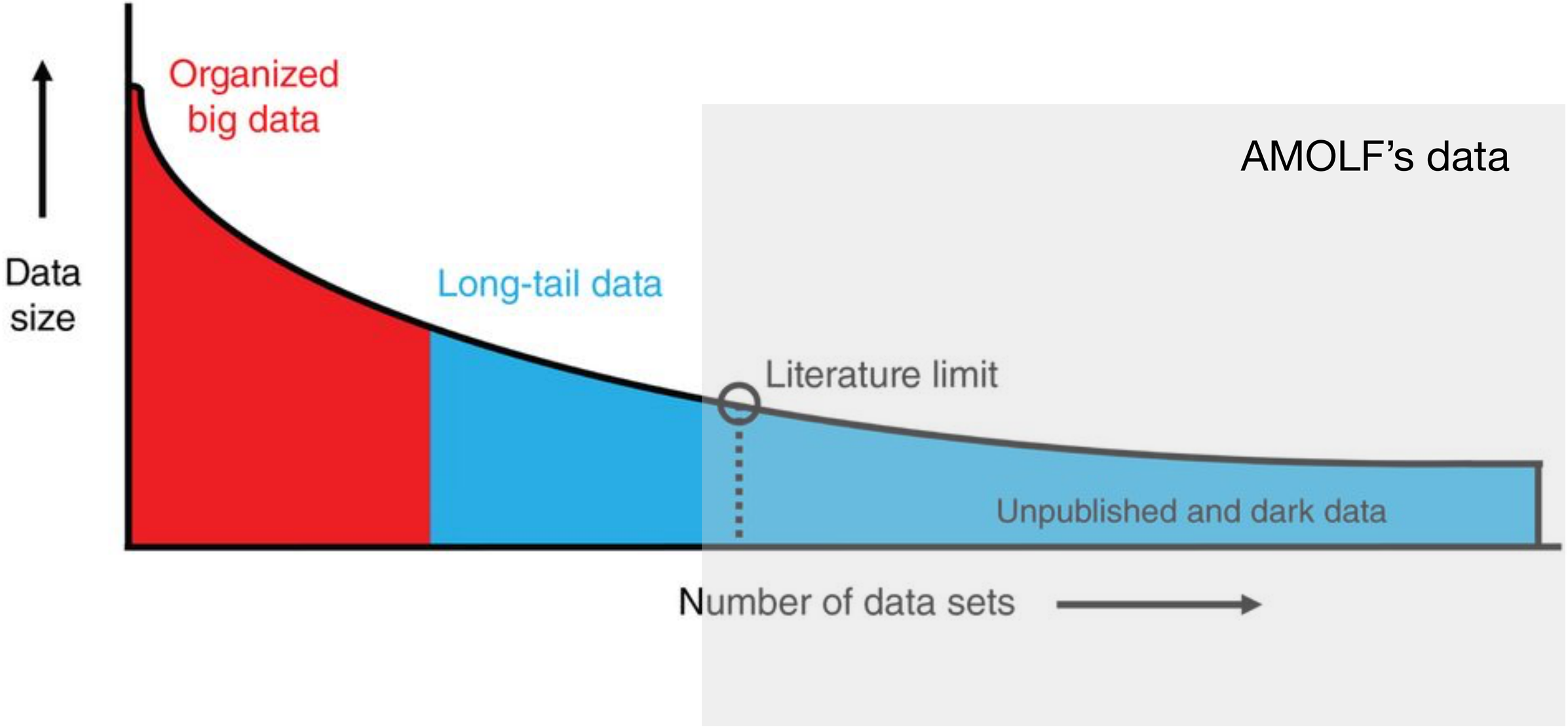
Voor de instituten is dit beleidskader opgesteld voor verantwoorde omgang met en beheer van onderzoeksdata met het oog op zowel de replicatie van het onderzoek als op het hergebruik van de onderzoeksdata voor nieuw onderzoek. De instituten werken dit kader verder uit.

The institutes will further elaborate this framework

Initial reactions

- We are scientists, we should spend our time on doing science, not bookkeeping
- This will mean extra time and costs: where will the money come from
- Nobody will want to see our data anyway

What is the value of our data?




WE ARE NOT
ENEMIES OF OPEN
DATA !

Pragmatism prevails

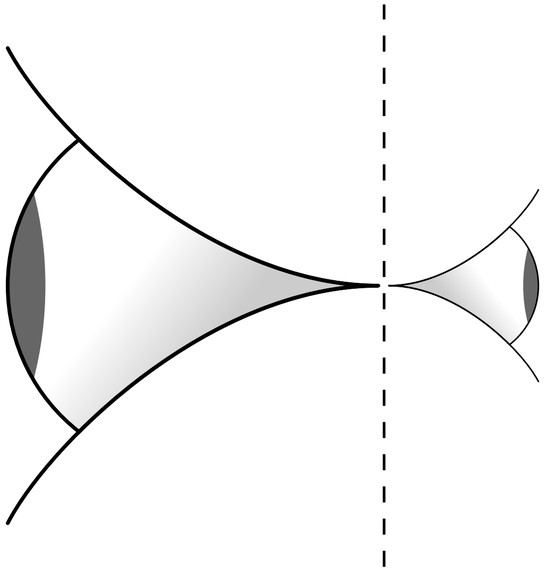
Ask not what you can do for datamanagement …
… but ask what datamanagement can do for you!

Create intrinsic value

How can we improve our own internal processes
with respect to data such that we ourselves profit?

Initial focus

Internal



External

Findable

Accesible

Interoperable

Reusable

3 goals

1. Make sure that datamanagement is seen as an integral part of doing research and part of the research planning process

2. Make sure that we have the proper infrastructure and tools to store and process data

3. Develop a strategy for the implementation of replication packages

Goal 1: embedding in organisation

Research logbooks

- Looked at many electronic options: no silver bullet
- Opted for more expensive paper logbooks, with page numbering, and a tracking system

Datamanagement plans

- Implement at group level (diversity)
- Create hierarchical templates to avoid redundant efforts

Goal 2: acquisition and storage

A tough one !

- Storage situation strained as it is:
  ~ 250 TB increasing by ~10 TB a year
- Many different types of data: from zillions of small text files to huge hi-res video files
- Hard to separate the temporary from the permanent
- Plethora of different acquisition systems/ formats, proprietary/home-grown

Goal 3: replication packages

Concrete dot on the horizon of datamanagement

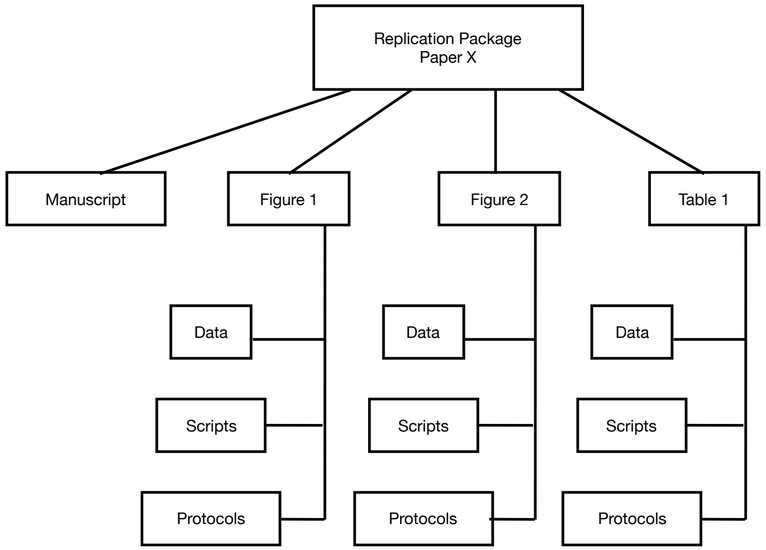
The goal of a replication package is to provide a ***minimal yet as complete as possible*** set of information by which an interested third party could ***in principle*** independently replicate the results of a paper ***published in a peer reviewed journal***. This material will typically include data, protocols, and analysis scripts. This material will be organized in a logical manner and provided with adequate metadata. In implementing replication packages we will be inspired by, and where feasible adhere to, the FAIR Data Principles.

Strategy

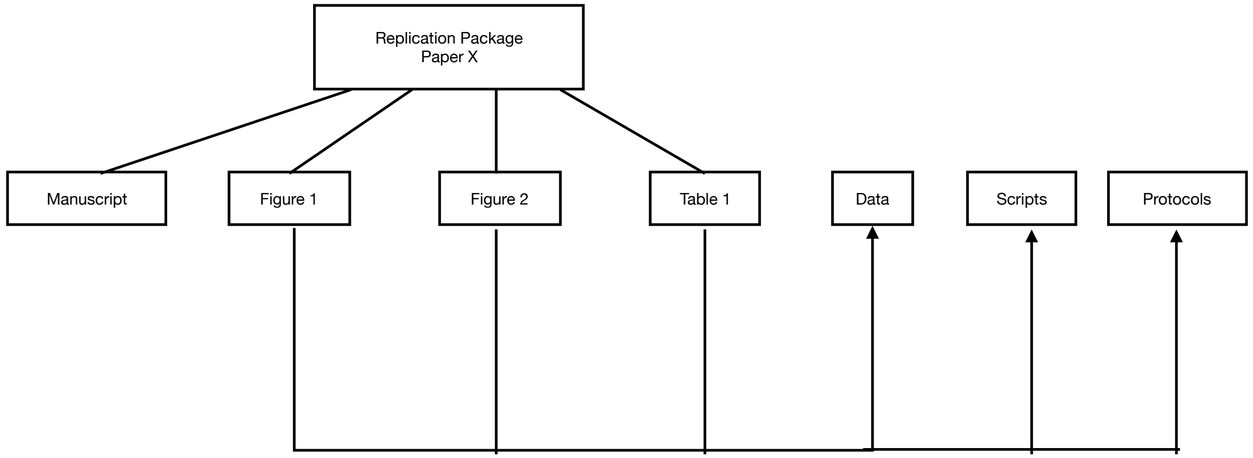
- Start light: minimal upfront requirements
- Take some time (~ year, ~ 100 papers) to develop best practices
- Explore open access only in a later stage once the process has crystallised

Replication package lite

Paper centric



Asset centric



Example

Quantification of Ion Migration in $CH_3NH_3PbI_3$ Perovskite Solar Cells by Transient Capacitance Measurements

Moritz H. Futscher[1], Ju Min Lee[1], Tianyi Wang[1], Azhar Fakharuddin[2], Lukas Schmidt-Mende[2] and Bruno Ehrler[1]

1. Center for Nanophotonics, AMOLF, Science Park 104, 1098 XG Amsterdam, The Netherlands
2. Department of Physics, University of Konstanz, Universitätsstraße 10, 78457 Konstanz, Germany

arXiv.org > cond-mat > arXiv:1801.08519

```
Futscher_Ehrler_Ion\ migration/
├── Figures
│   ├── Figure\ 1
│   │   ├── J(V)
│   │   │   ├── 180K
│   │   │   ├── 210K
│   │   │   ├── 240K
│   │   │   ├── 270K
│   │   │   ├── 300K
│   │   │   └── 330K
│   │   └── SEM
│   ├── Figure\ 2
│   │   ├── AFM
│   │   ├── C(V)
│   │   │   ├── 180K
│   │   │   └── 300K
│   │   └── Impedance
│   │       ├── 180K
│   │       ├── 210K
│   │       ├── 240K
│   │       ├── 270K
│   │       ├── 300K
│   │       └── 330K
│   ├── Figure\ 3
│   ├── Figure\ 4
│   │   └── TID
│   ├── Figure\ 5
│   └── TOC
├── Manuscript
└── Supplementary\ Figures
    ├── Figure\ S1
    │   └── J(V)
    │       ├── NiOx
    │       └── TiO2
    ├── Figure\ S10
    │   └── TID
    ├── Figure\ S2
    ├── Figure\ S3
    ├── Figure\ S4
    ├── Figure\ S5
    │   └── J(V)
    │       ├── 180K
    │       ├── 210K
    │       ├── 240K
    │       ├── 270K
    │       ├── 300K
    │       └── 330K
    ├── Figure\ S6
    │   └── Impedance
    │       ├── 180K
    │       ├── 210K
    │       ├── 240K
    │       ├── 270K
    │       ├── 300K
    │       └── 330K
    ├── Figure\ S7
    │   └── C(t)
    ├── Figure\ S8
    │   ├── Impedance
    │   │   ├── After\ first\ run
    │   │   │   └── 300K
    │   │   ├── After\ second\ run
    │   │   │   └── 300K
    │   │   └── Before
    │   │       └── 300K
    │   └── J(V)
    │       ├── After\ first\ run
    │       │   ├── 180K
    │       │   └── 300K
    │       ├── After\ second\ run
    │       │   ├── 180K
    │       │   └── 300K
    │       └── Before
    │           ├── 180K
    │           └── 300K
    ├── Figure\ S9
    │   └── TID
```

J-V characteristics at different temperatures

SEM images of samples

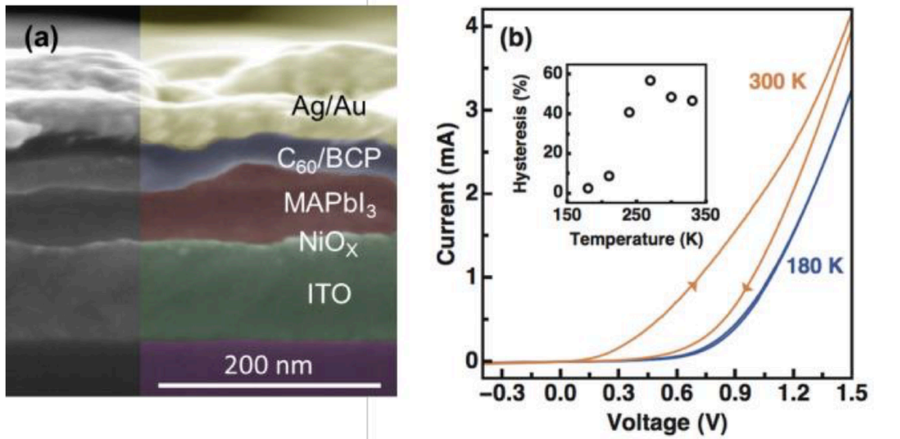


Figure 1

Moderate size: 70 MB

Open Issues

- Collaborations with third parties
- Determine the point of “no return” when package is “frozen”
- Dealing with errata
- Copyright issues (e.g. figures) upon open access

Take home messages

- Datamanagement plans need consensus and participation to be successful
- Commitment can be obtained if the internal use-case is strong
- It is not just a technical thing: the whole data workflow needs to be considered
- There are no one-size fits all silver bullet solutions
- Don’t underestimate the costs in terms of effort and hard-cash


Datamanagement paradise beckons
... but we still have quite a ways to go!