Smaller Hadron and Nuclear Communities
Stefan Typel

PUNCH4NFDI
Open Data Workshop
February 11, 2021
Overview

- Types and Sizes of Data
- Storing and Safeguarding of Data
- FAIRness of Data
- Data Publication and Repositories
- Support for Data Management
- Expectations

special thanks to Hannah Elfner and Jan Mayer for their thoughts here: only my personal view, no specific projects
Types and Sizes of Data

- very diverse range of topics: nuclear/hadronic structure and reactions, nuclear astrophysics, plasma, laser and atomic physics
- heterogeneous types of data: numeric data (different encoding schemes), software, handwritten notes/drawings, notebooks, logbooks, photos, texts, ...
- experiments
  - huge amount of raw data (typically 1 TB per experiment)
  - much smaller size after analysis
  - large archive of data from older experiments (several hundred TB)
- theory
  - often insignificant size
  - larger output from simulations (several 10 to 100 GB per run)
- data size small as compared to 'big' communities, but expected to increase
Storing and Safeguarding of Data

- raw data
  - mostly on local servers of individual institutes
  - often only short-term preservation
  - theory/simulations: reliance on reproducibility with codes and workflows

- analyzed data
  - local storage (sometimes accessible through web pages)
  - publication in journals (main article, supplementary material)
  - repositories (individual communities, university owned, special institutions)

- responsibilities
  - group leaders, PIs, system administrators
  - often no control mechanism

- long-term maintenance big unresolved issue

- meet requirements by funding agencies
FAIRness of Data

- Findable, Accessible, Interoperable and Reuseable
  - no systematic identification (DOI, ...)
  - very specific modes of access, if at all possible
    (often only by direct contact to authors)
  - heterogeneous systems of storage/hardware/software
  - only partly formalized meta data
  - numerical data, software, description of equipment at different places

- software
  - not all codes publicly available, openness vs. individual ownership
  - evolution and versioning, git repositories

- older data
  - F & A: enormous efforts required
  - I & R: unrealistic
Data Publication and Repositories

- largest part of data remain at institutes, mostly publication of highly processed data only
- rules of (international) collaborations to be followed
- types of repositories
  - local topical collections of particular data (operated by individuals)
  - university provided general repositories (sometimes not suitable)
  - special institutions (IAEA, NNDC@BNL, . . . , partly outdated)
- access to repositories
  - often no automated queries possible
  - no general standards
Support for Data Management

- data management relies on available equipment, financial/human resources
- operation of local systems/daily tasks
  - local IT people, postdocs, doctoral, master students
  - lack of specialists in data management
- sometimes involvement of university computing centers using centralized services
  - archiving, publication, citing, preparation of data management plans
- software tools
  - local installations of applications for sharing/synchronizing data/codes
  - use of commercial products (Github, Overleaf, ...)
- enforcement of policies by group heads (how seriously considered?)
- data management more and more integral part of specific projects (collaborative research centers, ...)

February 11, 2021 | PUNCH4NFDI Open Data Workshop | Stefan Typel | 7
Expectations

- development of reasonable and realizable methods for data management with limited resources
- unlocking of so far inaccessible data
- realisation of FAIR principles
- unification of software and tools for easy access, storing/sharing of data and data analysis (e.g. Bayesian statistical analysis)
- practical guides for users
- education of people on all levels
- raising awareness for data management

⇒ many challenges ahead of us
Thank you for your attention!