USQCD data management strategy and planning

R.G. Edwards
Jefferson Lab

J.N. Simone
Fermilab

April 27, 2019

Motivatation for this presentation

- Discuss the formulation of a data management strategy for USQCD.
- Seek your comments, and for you to provide feedback on the process.
- We would be happy to have you contribute your expertise to this effort!

Strategies and plans

- A data management strategy describes a coherent set of principles and objectives for managing data over all phases of their lifecycles: pre-planning, data production, distribution, active usage, and long-term preservation or eventual removal.
- A data management plan (DMP) describes the coordinated actions needed to accomplish the strategic goals.

U.S. policy: Who needs a plan?

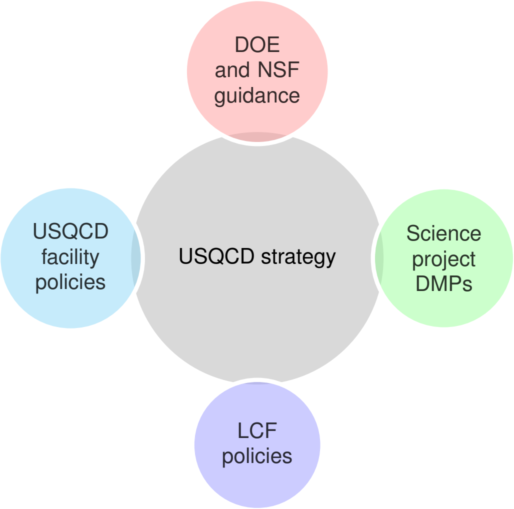
Policy on U.S. government sponsored research [DMPtool]:

> *In February of 2013, the White House Office of Science and Technology Policy (OSTP) issued a memorandum directing Federal agencies that provide significant research funding to develop a plan to expand public access to research. Among other requirements, the plans must: "Ensure that all extramural researchers receiving Federal grants and contracts for scientific research and intramural researchers develop data management plans, as appropriate, describing how they will provide for long-term preservation of, and access to, scientific data in digital formats resulting from federally funded research, or explaining why long-term preservation and access cannot be justified"*
>
> *The National Science Foundation (NSF) requires a 2-page plan as part of the funding proposal process. Most or all US Federally funded grants will eventually require some form of data management plan.*

Why planning is needed?

- The OSTP guidance does not suggest any additional data preservation funding, hence, careful planning is crucial.
- Future US directives on data: security, international access, scientific and technical competitiveness, . . . ?
- A DMP is a required for every DOE or NSF proposal. Proposals lacking a DMP are rejected without further consideration.
- European funding agencies have similar DMP requirements.
- USQCD needs a policy statement concerning preservation and access to data used in publications.
- Broader USQCD data policy benefits planning for preservation of data beyond scope of current project funding.
- Lack of a documented DM strategy was a finding in last yearly Project review.

USQCD strategy synthesis



USQCD strategy must consider:

- Conformance with DOE and NSF funding agency guidance.
- Differences among data policies at labs hosting resources used by USQCD.
- Policies in place at leadership computing facilities not under USQCD control.
- Needs and responsibilities of science projects allocated resources by USQCD.

How to organize documents?

Envision multiple documents encompassing USQCD strategy and other aspects rather than one document:

- A statement of USQCD strategic goals and policy objectives. Might include a meta-DMP for the whole project?
- Develop data policy documentation specific to each of BNL, FNAL, and JLab since policies do differ. See, e.g., FNAL policy statement
- A best practices guide to data handling for researchers, e.g. BEST PRACTICE FOR RESEARCHERS – UK Data Archive.
- Tools to create a DMP: a) add DMP template to new GoogleDoc project proposal worksheet, b) online tool (e.g. DMP tool) or c) Office templates, e.g., ▸ UVA DOE Generic to help each class A project develop an individual DMP. Having USQCD specific help is desirable. Example DMP ▸ HISQ Axial DMP

Strategic objectives

What follows is a first draft discussion of strategic objectives divided into three sections:

I. USQCD project-wide
II. Facility operations focused
III. Responsibilities of science projects and researchers

Defining and tuning these objectives will need input from the USQCD EC, SPC, site managers, and members.

I Project-wide objectives

1. A coherent and uniform data management strategy and policy will assist in effective allocation and utilization of computing resources project wide in order to maximize scientific productivity and impact.
2. The Scientific Program Committee (SPC) in consultation with USQCD facilites site managers will strive to best allocate available storage resources to the science projects.
3. USQCD is committed to ensuring research is reproducible and to providing open access to published data. The USQCD data management plan provides expectations for maintaining the provenance and integrity of published data and results.
4. USQCD will develop best practices for disseminating data used in publications.

I Project-wide objectives

5. QCD gauge ensembles are an import community resource that have been instrumental in enabling many physics projects. Gauge ensembles generated using USQCD resources will be available to the USQCD community.
6. QCD gauge ensembles generated by USQCD will remain an important scientific resource well beyond the current DOE Office of Science funded program(s). USQCD is committed to developing a long-term data preservation program for QCD gauge ensembles.
7. The record of scientific achievements enabled by USQCD produced gauge ensembles is invaluable. USQCD is working with inSPIRE-HEP and the DOE Office of Scientific and Technical Information (OSTI) to develop processes linking a publication to the data sets used in the publication via unique persistent Digital Object Identifiers (DOIs).
8. MILC asqtad ensembles are already citable by inSPIRE-HEP.

I Project-wide objectives: preservation of published data

- Scientific projects are expected to preserve the provenance and integrity of data and results that appear in publications. Typically, it is impractical to store all lattice data from intermediate stages leading to publications.
- Open access to data appearing in published plots and tables is provided by HEPdata. inSPIRE-HEP will cross reference papers and machine readable HEPdata files.
- Preserving data provenance, analysis codes, and intermediate data leading to published data is harder. Ideas for discussion:
- Preserve a (small) representative subset of output logs from production runs to document the data generation processes.
- Preserve reduced data (e.g. masses, matrix elements) leading to publications in an archival data format, e.g., HDF5, sqlite3, JSON.
- Jupyter notebook containing analysis codes and documentation.

II Facility objectives

1. Facilities hosting USQCD data have responsibility to implement the USQCD DM strategy in accordance with the DMP policies and plans of the respective laboratories.
2. Facilities will document data policies for users: BNL HEP Digital Data Management Policy, Fermilab data policy, and Jefferson Lab Data Management Plan.
3. The facilities maintain integrity and access to data over the life cycle of data sets. Data management plans from individual projects are necessary in order to document the data type, ownership and useful life span of data.
4. Active projects have highest priority in the allocation of resources. Inactive projects will be given a deadline to do housekeeping, after which facility managers will, at their discretion, migrate or delete the inactive data.

II Facility objectives

5. Facilities may provide projects short-term magnetic tape storage for the active life of the project and any extensions. Projects are expected to remove short-term data once a project becomes inactive.
6. Long-term magnetic tape storage may be available for community data or other important data. Recurring costs to the Project for long-term storage include a single migration of the data to newer tape technologies. Additional migrations may incur additional costs in addition to recurring costs.

III Responsibilities of researchers

1. A DMP is a useful to help researchers understand the costs and responsibilities with regard to their data. A DMP can help a project understand and better balance compute vs storage requirements and costs when developing workflows.
2. Class 'A' scientific projects are allocated a significant amount of scarce computing resources. Class A projects and are expected create, maintain, and execute their data management plan. A detailed DMP is optional for a Class 'B' and 'C' projects unless the project will involves managing exceptionally large quantities of data.
3. Each scientific project should identify a data manager. The data manager will make data management decisions on behalf of the project and they, with the project PI, will be the primary points of contact regarding data issues.

III Responsibilities of researchers

4. Adopt best practices for data organization, data storage formats and metadata markup. Poor choices often impacts both the productivity of the project and the performance of the host facility.
5. Individual researchers and each project are ultimately responsible for the integrity of their data. Use backup procedures such as replicating important data among geographically separated locations to prevent the catastrophic loss.
6. Each project is expected to abide by the data management policies of the facility hosting their data. A project seeking an exemption should immediately notify the facility and the Scientific Program Committee and provide justification for the exception.

III Responsibilities of researchers

7. Projects may request long-term magnetic tape storage for critical results or community data such as QCD gauge configurations. Projects should request long-term storage as part of their proposal and data management plan. The amount of storage and hosting institutions will be negotiated by the SPC, EC, and the USQCD facility managers.
8. Each project needs to have a plan for what is to happen to their data after the becomes inactive or ends. At the earliest stages of planning, a project should begin to develop a workable, cost effective plan for data preservation.
9. Not all data created by a project need be shared or preserved. The costs and benefits for doing so needs to be weighed during data management planning.
10. Projects seeking long-term data preservation should begin negotiating with suitable data hosting sites as soon as possible. Projects may also need to apply to the funding agencies or their host institutions for additional funds for data preservation.

Summary

1. We need your help, feedback and expertise!
2. Write USQCD strategy document before the summer Project review.
3. Complete BNL, Fermilab, and JLab facilities data management policy documentation.
4. Create a DMP template for use in the USQCD class A proposal process.
5. Draft a data best practices guide.
6. Be mindful of your data’s life cycle.

UVA DOE generic template

▸ BACK

Department of Energy (DOE)

[Replace Header with ‘Data Management Plan’ prior to submission]

Data types and sources

[Enter content here, and then remove the Guidance prior to submission]

Guidance:
A brief, high-level description of the data to be generated or used through the course of the proposed research and which of these are considered digital research data necessary to validate the research findings.

DMPs should provide a plan for making all research data displayed in publications resulting from the proposed research open, machine-readable, and digitally accessible to the public at the time of publication. This includes data that are displayed in charts, figures, images, etc. In addition, the underlying digital research data used to generate the displayed data should be made as accessible as possible to the public in accordance with the principles stated above. This requirement could be met by including the data as supplementary information to the published article, or through other means. The published article should indicate how these data can be accessed. The term digital data encompasses a wide variety of information stored in digital form including: experimental, observational, and simulation data; codes, software and algorithms; text; numeric information; images; video; audio; and associated metadata. It also encompasses information in a variety of different forms including raw, processed, and analyzed data, published and archived data. This statement focuses on digital research data, which are research data that can be stored digitally and accessed electronically. Research data are defined in regulation (2 CFR 200.315 (e), continuing the definition from 2 CFR 215 (OMB Circular A-110) as follows: “Research data is defined as the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. This 'recorded' material excludes physical objects (e.g., laboratory samples). Research data also do not include: (A) Trade secrets, commercial information, materials necessary to be held confidential by a researcher until they are published, or similar information which is protected under law; and (B) Personnel and medical information and similar information the disclosure of which would constitute a clearly unwarranted invasion of personal privacy, such as information that could be used to identify a particular person in a research study.”

Content and format

[Enter content here, and then remove the Guidance prior to submission]

Guidance:
A statement of plans for data and metadata content and format including, where applicable, a description of documentation plans, annotation of relevant software, and the rationale for the selection of appropriate standards. (Existing, accepted community standards should be used where possible. Where community standards are missing or inadequate, the DMP could propose alternate strategies that facilitate sharing, and should advise the sponsoring program of any need to develop or generalize standards.)

DMPs should reflect relevant standards and community best practices for data and metadata, and make use of community accepted repositories whenever practicable.

Sharing and preservation

[Enter content here, and then remove the Guidance prior to submission]

Guidance:
Data sharing means making data available to people other than those who have generated them. Examples of data sharing range from bilateral communications with colleagues, to providing free, unrestricted access to the public through, for example, a web-based platform. Data preservation means providing for the usability of data beyond the lifetime of the research activity that generated them.

- The section on sharing and preservation should include, when appropriate: the anticipated means for sharing and the rationale for any restrictions on who may access the data and under what conditions.
- a timeline for sharing and preservation that addresses both the minimum length of time the data will be available and any anticipated delay to data access after research findings are published.
- any special requirements for data sharing, for example, proprietary software needed to access or interpret data, applicable policies, provisions, and licenses for re-use and re-distribution, and for the production of

derivatives, including guidance for how data and data products should be cited.

- any resources and capabilities (equipment, connections, systems, software, expertise, etc.) requested in the research proposal that are needed to meet the stated goals for sharing and preservation. (This could reference the relevant section of the associated research proposal and budget request).
- cost/benefit considerations to support whether/where the data will be preserved after direct project funding ends and any plans for the transfer of responsibilities for sharing and preservation.
- whether, when, or under what conditions the management responsibility for the research data will be transferred to a third party (e.g. institutional, or community repository).
- any other future decision points regarding the management of the research data including plans to reevaluate the costs and benefits of data sharing and preservation.

DMPs should consult and reference available information about data management resources to be used in the course of the proposed research. In particular, DMPs that explicitly or implicitly commit data management resources at a facility beyond what is conventionally made available to approved users should be accompanied by written approval from that facility. In determining the resources available for data management at Office of Science User Facilities, researchers should consult the published description of data management resources and practices at that facility and reference it in the DMP.

Protection

[Enter content here, and then remove the Guidance prior to submission]

Guidance:
A statement of plans, where appropriate and necessary, to protect confidentiality, personal privacy, Personally Identifiable Information, and U.S. national, homeland, and economic security; recognize proprietary interests, business confidential information, and intellectual property rights; and avoid significant negative impact on innovation, and U.S. competitiveness.

DMPs must protect confidentiality, personal privacy, Personally Identifiable Information, and U.S. national, homeland, and economic security; recognize proprietary interests, business confidential information, and intellectual property rights; avoid significant negative impact on innovation, and U.S. competitiveness; and otherwise be consistent with all applicable laws, regulations, and DOE orders and policies. There is no requirement to share proprietary data. Personally Identifiable Information for proposals with Human Subjects Research (HSR), including research involving Personally Identifiable Information (PII), an appropriate research protocol will need to be approved by the appropriate DOE Institutional Review Board (IRB) or an external IRB with an approved federal wide assurance. Follow the instructions of the research solicitation to determine whether or not the data management aspects of this protocol should be included in the DMP. At a minimum, the DMP should acknowledge the type of HSR and/or PII involved and give a projected timeline for IRB approval. Information regarding DOE requirements for HSR and research involving PII, including how to obtain IRB approval, can be found at this link: DOE Human Subjects FAQ.

Rationale

[Enter content here, and then remove the Guidance prior to submission]

Guidance:
A discussion of the rationale or justification for the proposed data management plan including, for example, the potential impact of the data within the immediate field and in other fields, and any broader societal impact.

At a minimum, DMPs must describe how data sharing and preservation will enable validation of results, or how results could be validated if data are not shared or preserved.

Additional Guidance (Optional)

HISQ Axial ▸ BACK

THE NUCLEON AXIAL-VECTOR FORM FACTOR AT THE PHYSICAL POINT WITH THE HISQ

A Data Management Plan created using DMPTool

Creators: James Simone, James Simone

Affiliation: University of Chicago

Template: Department of Energy (DOE)

Project abstract:
We propose to continue our computation of the axial-vector form factor of the nucleon using the highly- improved staggered-quark (HISQ) action for both valence and sea quarks. We use the (2+1+1)-flavor HISQ ensembles generated at the physical point, combining lattice QCD calculations of the q 2 dependence with the z expansion to obtain a model-independent description of the shape. As a by-product, we will compute the axial charge g A directly at the physical point. We will test our approach with the shape of the vector form factor, which is constrained by high-statistics electron-scattering data. The project is well aligned with USQCD goals, because the axial-vector form factor is an important ingredient in quasielastic neutrino- nucleon scattering, which is the key signal process in neutrino-oscillation experiments at Fermilab. We request 203 kGPU-hours and 3 M Sky-core-hours at BNL or Fermilab; we also request 50 Tbyte disk space and 108 Tbyte tape storage. If we continue running at BNL and the tape storage has to be at Fermilab or JLab, some scratch space for staging to the tape robot is also needed. Using USQCD conversion factors, the total request is 1184 M Sky-core-hours.

Last modified: 02-25-2019

THE NUCLEON AXIAL-VECTOR FORM FACTOR AT THE PHYSICAL POINT WITH THE HISQ ENSEMBLES

1. DATA SHARING AND PRESERVATION

Data management plans should describe whether and how data generated in the course of the proposed research will be shared and preserved. If the plan is not to share and/or preserve certain data, then the plan must explain the basis of the decision (for example, cost/benefit considerations, other parameters of feasibility, scientific appropriateness, or limitations discussed in #4). At a minimum, DMPs must describe how data sharing and preservation will enable validation of results, or how results could be validated if data are not shared or preserved.

For this project there are three classes of data that need to be shared and/or preserved:

1. Staggered corner-wall propagators on the second generation a=0.15, 0.12 and 0.09fm MILC ensembles, including metadata. This data totals 108TBs.
2. Unaveraged two- and three-point correlator data on the above ensembles, including metadata.
3. Plotted data used in figures and tables for publications using the data given in 1) or 2).

Preservation:

1. Propagator data will be kept on tape storage at FNAL/BNL for the duration of the project, or for as long as USQCD deems fit, whichever is the shorter. These propagators are preserved because it is significantly more expensive to regenerate new propagators for the final physics goals (the nucleon axial form factor) than it is to keep them. These are stored in standard MILC format.
2. The unaveraged correlator data - including metadata - will be kept in sqlite3 databases (a standard format accepted as future proof by office of science). This data will stored on disk where USQCD gives storage as part of our allocation for the duration of the project. After the project has met it's goals and finished, the data will be kept an additional 3-5 years on disk. If funding is not allocated for this storage, then personal harddrives will be used to store the data.
3. The plotted data will be submitted to HEPdata.net and or kept in sqlite3 databases and will follow the same preservation plan as data in 2.

Sharing:

1. We already encourage projects to use our staggered propagators if possible. If anyone wants these propagators they are encouraged to contact us. For projects that are not in direct competition with our physics goals, the propagators will be made available instantly. For competing projects, the propagators will be made available as soon as our physics goals are met.
2. The unaveraged correlator data in sqlite3 databases - including metadata - will be made available upon request as soon as the time of publication. The publication will indicate this.
3. Identical to data in 2).

Management:

- The management of the data will fall under the remit of the PI of this grant, and if the PI changes then the subsequent PI will take over management also.

2. DATA USED IN PUBLICATIONS

Data management plans should provide a plan for making all research data displayed in publications resulting from the proposed research open, machine-readable, and digitally accessible to the public at the time of publication. This includes data that are displayed in charts, figures, images, etc. In addition, the underlying digital research data used to generate the displayed data should be made as accessible as possible to the public in accordance with the Principles published in the DOE Policy for Digital Research Data Management. The published article should indicate how these data can be accessed.

Addressed in Section 1.

- All data will be made publically available at the time of publication, if not before.

- All data will be in standard format: 1) propagators in MILC format; 2) correlator and results/figures/tables in sqlite3 database.

3. DATA MANAGEMENT RESOURCES

Data management plans should consult and reference available information about data management resources to be used in the course of the proposed research. In particular, DMPs tht explicitly or implicitly commit data management resources at a facility beyond what is conventionally made available to approved users should be accompanied by written approval from that facility. In determining the resources available for data management at DOE Scientific User Facilities, researchers should consult the published description of data management resources and practices at that facility and reference it in the DMP. Information about other Office of Science facilities can be found in the additional guidance from the sponsoring program.

All resources (FNAL/BNL tape and disk storage) are under the control the USQCD executive committee and so we do not require additonal permissions to hold propagators or databases at such facilities.

4. CONFIDENTIALITY, SECURITY AND RIGHTS

Data management plans must protect confidentiality, personal privacy, Personally Identifiable Information and U.S. national, homeland, and economic security; recognize proprietary interests, business confidential information, and intellectual property rights; avoid significant negative impact on innovation and U.S. competitiveness; and otherwise be consistent with all applicable laws, regulations, agreement terms and conditions, and DOE orders and policies. There is no requirement to share proprietary data.

No data breaches any confidentiality or pose any security issues for the U.S., or any associated facility.