Belmont Forum Data Accessibility Statement and Policy

Status: Approved by Plenary, 18 October 2018. Will be reviewed regularly and revised as needed, in consultation with the science publishing community.

Underlying Rationale

In 2015, the Belmont Forum adopted the Open Data Policy and Principles . The e-Infrastructures & Data Management Project is designed to support the operationalization of this policy and has identified the Data Publishing Policy Project (DP3) as a key activity towards this objective.

Recently, approaches to open and FAIR (Findable, Accessible, Interoperable and Re-usable[1]) data have begun to converge across the funding, research and publishing communities, although there may be some differences in approaches and goals. Among the similarities is the drive towards encouraging high quality published research that, together with its underlying data, is reproducible and potentially re-usable by a range of communities. Toward that end, both funders and publishers have opportunities to communicate with researchers through policy and guidance at various points during the research cycle. However, they have rarely attempted to coordinate these efforts.

The DP3 has been explicitly co-developed by the Belmont Forum and science publishing community[2] to complement the Data and Digital Outputs Management Plan (DDOMP) requirements being incorporated into all future Belmont Forum Collaborative Research Actions (CRAs) or funding calls. The DDOMP is designed to provide Belmont Forum grantees a framework for practicing good management of transnational and transdisciplinary data throughout their research projects, from the pre-proposal stage to its conclusion. By linking the language and expectations of the DDOMP with the DP3 initiative, funders and publishers will articulate a coherent set of data and research outputs management expectations for researchers, with the ultimate result of improved sharing and re-use of research data[3]

The DP3 is concentrating chiefly on delivering a set of template Data Accessibility Statements[4] (DASs) for guiding Belmont Forum grantees when publishing their research results. The DAS is included as part of a journal article and articulates which data underlie a paper, where the data are available and under what conditions they can be accessed.

The DAS encourages researchers to plan for the longevity, reusability, and stability of the data attached to their research publications and results. Access to data promotes reproducibility, prevents fraud and thereby builds trust in the research outcomes based on those data amongst decision- and policy-makers, in addition to the wider scientific community and the general public. Even when the underlying dataset cannot be open (for example, for confidentiality, conservation or security reasons), good machine readable metadata and information about the data's licensing restrictions are important. Furthermore, by supporting data citation, the DAS also provides opportunities to credit the data collectors and curators.

---

[1] Link to FAIR Data Principles information here: https://www.force11.org/group/fairgroup/fairprinciples

[2] The DP3 Advisory Group consists of representatives from PLOS, Hindawi, Elsevier, Springer Nature, Digital Science and CODATA, as well as members of the Belmont Forum e-Infrastructures & Data Management Project and Data Liaisons from Belmont Forum member agencies.

[3] The primary purpose of the DAS is to provide a high quality record of the relevant research data robustly linked with the primary journal article. The DDOMP, however, is concerned with the management and stewardship of all the data and digital outputs associated with a project beyond the project's actual duration and regardless of whether they are associated with a publication or not.

[4] Sometimes called 'Data Availability Statements'.

DAS Requirements

1. The DAS is to be situated in front of any paywall so that it can be read by humans and machines without the need to pay any fees.
2. The DAS will include the following:
    - **Confirmation** the data underlying a publication exists
    - Information on **where** the data can be found[5]
    - **Persistent identifiers** where available, including DOIs, Accession Numbers, IGSNs, other domain-recognised PIDs (such as those used in Astronomy), and standard terms for repositories/institutions (such as GRID/ISNI/OrgRef). We are aware that these systems are in the process of developing so Grant IDs, for instance, may become compulsory at a later point in time[6]
    - **Licensing** restrictions and **access** requirements (e.g., registration and fees)

NB: A specific DAS may consist of several statements relating to one or more datasets.

Workflow

A typical order of events would be for the grantee/PI to have a DAS template at the start of the research project with the advice that they should populate it (referencing the DDOMP) to accompany their article when submitting to journals.

It is anticipated that more work will need to be done to ensure this workflow is sufficiently connected throughout the research cycle and embedded within the messaging and practices of the CRAs, funding bodies and other BF stakeholders.

[5] Including (1) reference to the specific data underpinning the reproducibility of the relevant research article (subset) and to the overall dataset(s) relating to the project and (2) ideally the data landing page will include current contact details, as well as any charges for accessing.

[6] The name of Project and Granting Agency (until Grant Identifiers mature)

DAS Section by Section

Below is a mapping of the DAS requirements, together with explanatory characteristics and initial examples.

Requirement	**Characteristic(s)**	**Example(s)**
Confirmation the data do (or do not) exist	Where available, use persistent identifiers, such as DOIs	Crystal structures are accessible, subject to registration, from "My University" [grid.423328.c][7] at **https://doi.org/10.15125/010203**
	This covers **re-analyses** of existing data	This study was a re-analysis of existing data that are available from XXXX [grid.23345] at https://doi.org/10.15125/12345. Further documentation about data processing are available subject to registration from the XXXX [grid.317] at https://doi.org/10.15125/12345.
	Situations where there are no new data	**No new data** were created during this study.[8]
	Where there is **a large number of datasets** that need to be listed elsewhere from the actual DAS.	The study brought together existing data obtained upon request and subject to licence restrictions from a number of different sources. Full details on how these data were obtained, including individual persistent identifiers, workflow analyses and processing, are available in the documentation available at https://doi.org/10.15125/12345.
Information on where the data underlying the article can be found	Ideally the data landing page will include current contact details, as well as any charges for accessing.	Microscopy images are openly accessible, using a CC-0 licence from XXXX [grid.466587.e] at https://doi.org/10.15125/01423 **NB Specific contact details and charges for accessing should not be included in**

[7] Global Research Identifier Database (GRID) is part of the Digital Science portfolio and downloadable using a Creative Commons Public Domain 1.0 licence. Other research organisational identifier systems are available or in development.

[8] This is very unusual. We would expect to see it in the Belmont Forum context only for a speculative or opinion article (which may not require a DAS in any case).

		the DAS itself as they are likely to change over time.
As well as reference to the specific data underpinning the reproducibility of the relevant research article (subset) and to the overall dataset(s) relating to the project	This is so that the connections between datasets relating to the same project can be made (easily) explicit. There is still some work to be done on how the relative timing and relationships between specific (to particular articles) datasets and general (relating to overall projects) datasets will be mapped.	Information **on the full suite of datasets** relating to the **"Mapping Rivers and Impacts" Program** can be accessed at https://doi.org/10.15125/12345
Persistent identifiers where available including **DOIs**, **Accession Numbers**, **IGSN**s. We are aware that these systems are in the process of developing so **Grant ID**s, for instance, may become compulsory at a later point in time	Persistent identifiers enable connections to be made between the entities themselves rather than simply pointing to internet locations. Ultimately, it is anticipated that incorporating PIDs into research workflows will support automation of services.	This paper contains several samples identified by IGSN, one of them is IGSN: SSH000SUA. Information about this sample can be obtained by resolving the IGSN by adding the URL of the resolver before the IGSN: **http://igsn.org/SSH000SUA.**
Standard terms for repositories/institutions (such as **GRID/ISNI/OrgRef**)	This is to minimise confusion between institutions and maximise opportunities to make useful connections via automated search terms. Should decide which system the Belmont Forum will use.	Microscopy images are openly accessible, using a CC-0 licence from **XXXX University [grid.466587.e]** at https://doi.org/10.15125/01423
Licensing restrictions	Currently, the Belmont Forum isn't in a position to mandate a specific licence, or suite of licences. However, policy wording such as 'a CC-BY licence or its equivalent' could be workable.	Microscopy images are openly accessible, using a **CC-0 licence** from XXXX University [grid.466587.e] at https://doi.org/10.15125/01423

Access requirements (e.g., ethics, environmental concerns[9],[10], privacy, registration and fees)	Care needs to be taken to distinguish between DAS level information (which should have long-term accuracy) and more transitory features (that should be included on the landing page or elsewhere)	Due to the fact that **human subjects** are involved[11], supporting data cannot be made openly accessible. Further information about the data and conditions for access can be found at the University of XXX [grid.7340.0] data archive: https://doi.org/10.15125/1234
The name of Project and Granting Agency/Agencies (until Grant Identifiers mature). Use standard wording as provided by the CRA developers.	Due to its anticipation that PID uses will continue to evolve, this transitional requirement will eventually be replaced by Project and/or Granting Agency IDs.	The underlying data relating to this article were funded by the **National Science Foundation, Japanese Technology Fund, UK Research Fund and National Taiwan Research** within the **Belmont Forum "Mapping Rivers and Impacts" Program**.[12]

EXCLUSIONS

The DAS *should not* include the following:

NB: These factors may still be permissible when the research in question was not funded by the Belmont Forum (for instance, if the research was unfunded, or commercially funded).

1. URLs in place of persistent identifiers (where persistent identifiers are available)
2. The statement "data are in the supplementary information/supporting material section" (this is an inherently unstable solution so the data in question would always be better served by being lodged in an approved repository)
3. The statement "all data are in the results section" (this may be behind a paywall and it also ignores the requirement to reference the full project dataset in the DAS)
4. Referring simply to the reference list in general (this may be behind a paywall and in any case, the DAS provides more useful information about which data actually underpin the research)
5. Keeping data closed because of commercial restrictions (unless there has been a specific agreement with the funders, this is not an option for Belmont Forum funded research)

[9] Consistent with the Environmental Data Initiative policy: https://environmentaldatainitiative.org/data/edi-data-policy/

[10] Although the Natural Environment Research Council Data Policy document (authored by Mark Thorley and Sarah Callaghan) only envisages potential embargo periods to allow first publication opportunities to collectors and does not mention environmental sensitivities per se: https://nerc.ukri.org/research/sites/data/policy/datapolicy-guidance/

[11] From the DDOMP language: 'Belmont Forum policy is that the data should be as open as possible to commercial and non-commercial users, though with managed access where appropriate and necessary, for example, if there are sensitive data involving human subjects.' There may also be environmental, geopolitical or other reasons for not making data immediately open.

[12] Consider including information on project duration, PIs, etc on the landing page but not the DAS itself.

Other Key Issues

1. Needs to fit with other community initiatives such as: https://jats4r.org/data-availability-statements
2. The distinction in function between the DAS and Data Citation.
    - The DAS provides a robust link between primary research and its underlying dataset that is in front of the paywall. This is also an opportunity to provide additional information above that included in a standard citation that will support the article's reproducibility, as well as the potential for the data to be re-used meaningfully.
    - Data should also be cited in the references. This has implications for human and machine readability, take up by indexers, and connects with the Data Citation Implementation Pilot.
3. Repositories. The Belmont Forum Open Data Policy and Principles require repositories to be trustworthy. Wording such as 'selected repositories should fulfill research community standards' would be sufficient in the first instance, but this should be revisited through subsequent iterations. In the longer term, Belmont Forum-funded projects may
    - align with the Go-FAIR initiative that is working with some publishers on specifying criteria that can be used to filter 'recommended repositories' for publishers
    - adopt the Enabling FAIR Data language and requirements for 'FAIR-compliant Repositories' (see Appendix C) as capabilities emerge
    - require compliance with the CoreTrustSeal standard.

DAS Policy Does Not Cover

1. Non-digital data that cannot be digitised.[13]
2. Policing and Compliance.
    a. Currently, BF funders are at different stages in their interest and ability to monitor and enforce. As the DAS project (and others) mature, this may develop into a service that publishers can provide.
    b. The costs and funding of compliance with DAS and DDOMP requirements.
    c. Timing for evaluation
    d. Who should perform evaluations?
3. URL vs DOI – many policies are on websites; is it possible to have a policy DOI, which would make the policy that was applicable at the time of publication persistent even with changes over time?
4. Which licence to use.
5. When in the course of the research process the data need to be made accessible (as in, before or after publication, or at the completion of a specific Work Package or at the end of the whole project, etc).

[13] Examples of non-digital that can be digitised include analogue maps, tapes and images. Non-digital data that cannot be digitised includes physical samples.

Glossary of Terms

1. Research Data -the DAS takes a broad definition approach. ''Research Data" refers to any digital research output that would realistically be required in order for the relevant article to be adequately validated. Consequently, the DAS template employs the DDOMP definition of 'data and digital outputs, which include, but are not limited to:
    a. Quantitative and qualitative digital information and objects created during or reused in research activities such as experiments, analyses, surveys, interviews, measurements, instrumentation, observations, video, audio, and computer simulations;
    b. All metadata describing the data and digital outputs, their acquisition (including model description and related metadata for simulations and workflows), and other details for the use and the reuse of the data;
    c. Secondary data resulting from data reduction, transformation, analyses, and results, together with the associated code, software, workflows, and provenance information;
    d. Stakeholder-oriented digital outputs such as maps (including GIS layers), decision support tools, tutorials, videos, local language resources, lesson plans, curricula, policy memos, and whitepapers; and
    e. Descriptions of, and metadata relating to, physical samples connected with the CRA - but not the actual physical samples.

2. Gradations of 'Accessibility' within statements
    a. Openly accessible
    b. Subject to registration
    c. Not openly accessible but conditions for accessing are included

Appendix A: Review process

Who	What	When	Done/Remarks
Bob Samors/Fiona Murphy	First Draft	17-2 May 2018	Circulated to DP3 Group
Data Publishing Advisory Group (including BF Data Liaisons)	Second Draft	2-25 May 2018	Further updates made. Possible CSIRO project developing
Belmont Forum Steering Committee	Second Draft	4 June 2018 October 2018	
Enabling FAIR Data/Projects/Other Initiatives (Shelley Stall, AGU; David Carr, The Wellcome Trust; Josh Brown, ORCID)	Various Drafts	May-August 2018	Iterations and updates incorporated
Belmont Forum Plenary	Final Draft	October 2018	

Appendix B: Contributors/Community Feedback Events

Attendees of Belmont Forum Data Publishing Policy Workshop (June 2017)
Members of Advisory Group monthly teleconferences (2018)
Data Liaisons Telecons (2018)
Researcher to Reader Workshop (February 2018): librarians and publishers
Enabling FAIR Data Initiative Potsdam Workshop (March 2018)
Data Policy Standardisation & Implementation Interest Group breakout session at RDA Berlin Plenary (March 2018)
OECD Global Science Forum Presentation (April 2018)
Consultations with other projects and experts, such as Shelley Stall of the AGU/Enabling FAIR Data; David Carr/Wellcome Trust; Josh Brown/ORCID

Appendix C: FAIR-compliant Repositories

This table is adapted from the list of services and benefits that will be expected from FAIR-compliant repositories in due course according to the Enabling FAIR Data Project.

Services Provided:	**Benefits:**
Metadata support: Repository, Datasets, Citation	Supports Discovery, Understanding, Reuse
Persistent identifiers	Supports Data Citation and Credit for Data and Reuse
Citation / Landing page compliance	Supports Best Practices and Common Experience for Researchers
Licensing policies (data and software)	Supports reuse of data and software.
Common list of approved FAIR-compliant repositories	Supports researchers locating compliant repositories.
Third-party validation of repository capability	Certification validating that many of the elements described above are implemented correctly in the repository (e.g. CoreTrustSeal, DataOne Member Node).