

Towards harmonised principles for data preservation, re-use and (open) access at the LHC

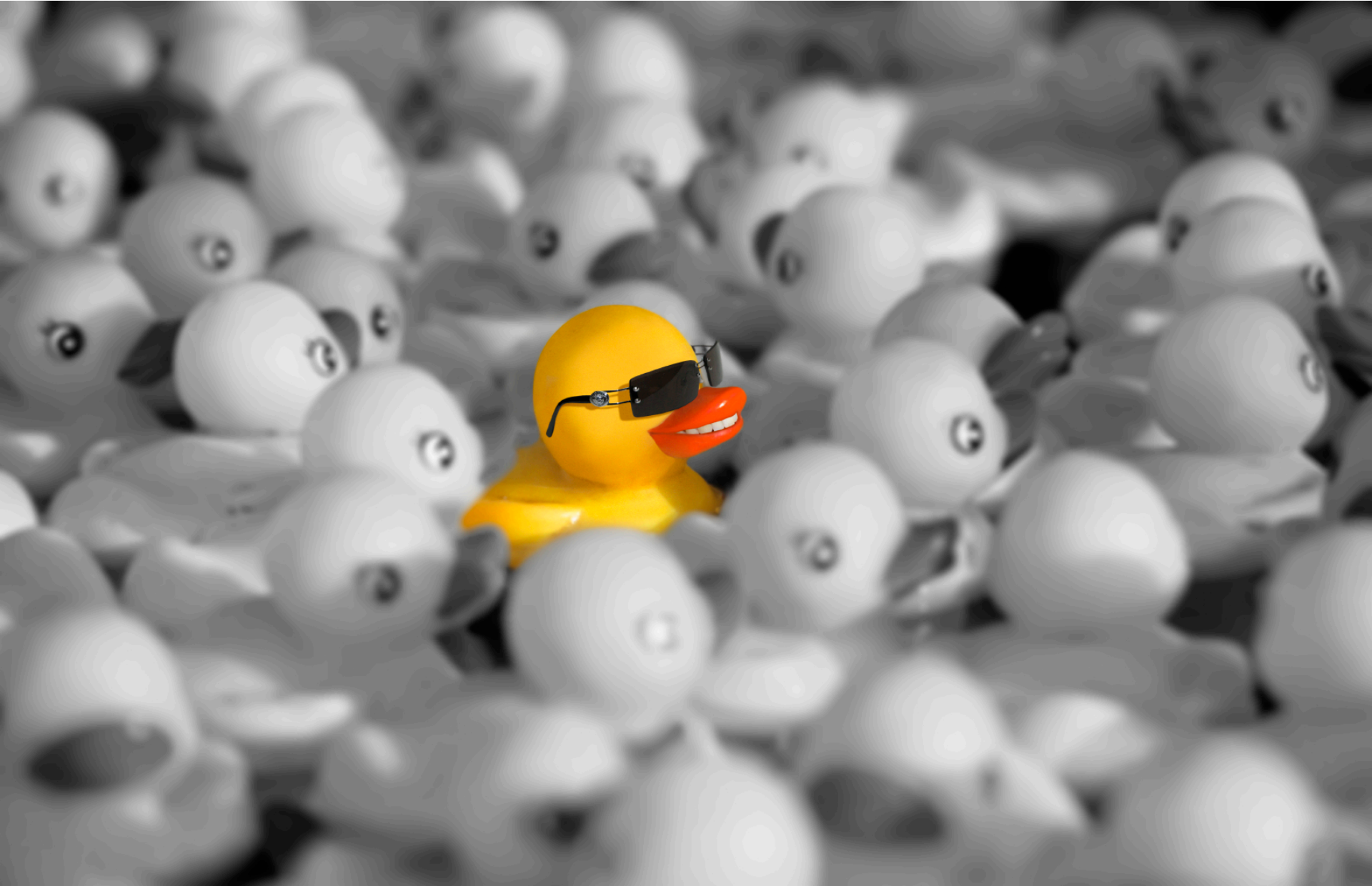


Sünje Dallmeier-Tiessen, Salvatore Mele
CERN – Scientific Information Service/Open Access
DPHEP@Marseille November 18th 2012

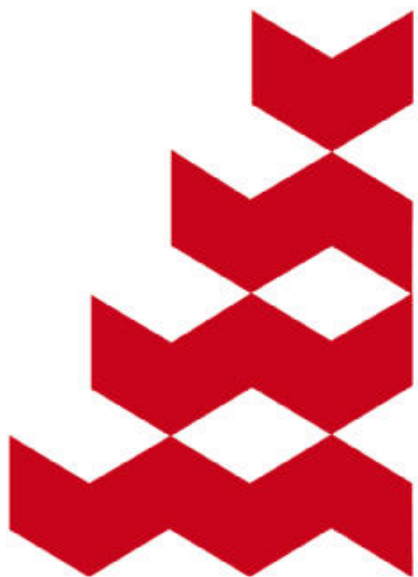




Re-use matters!

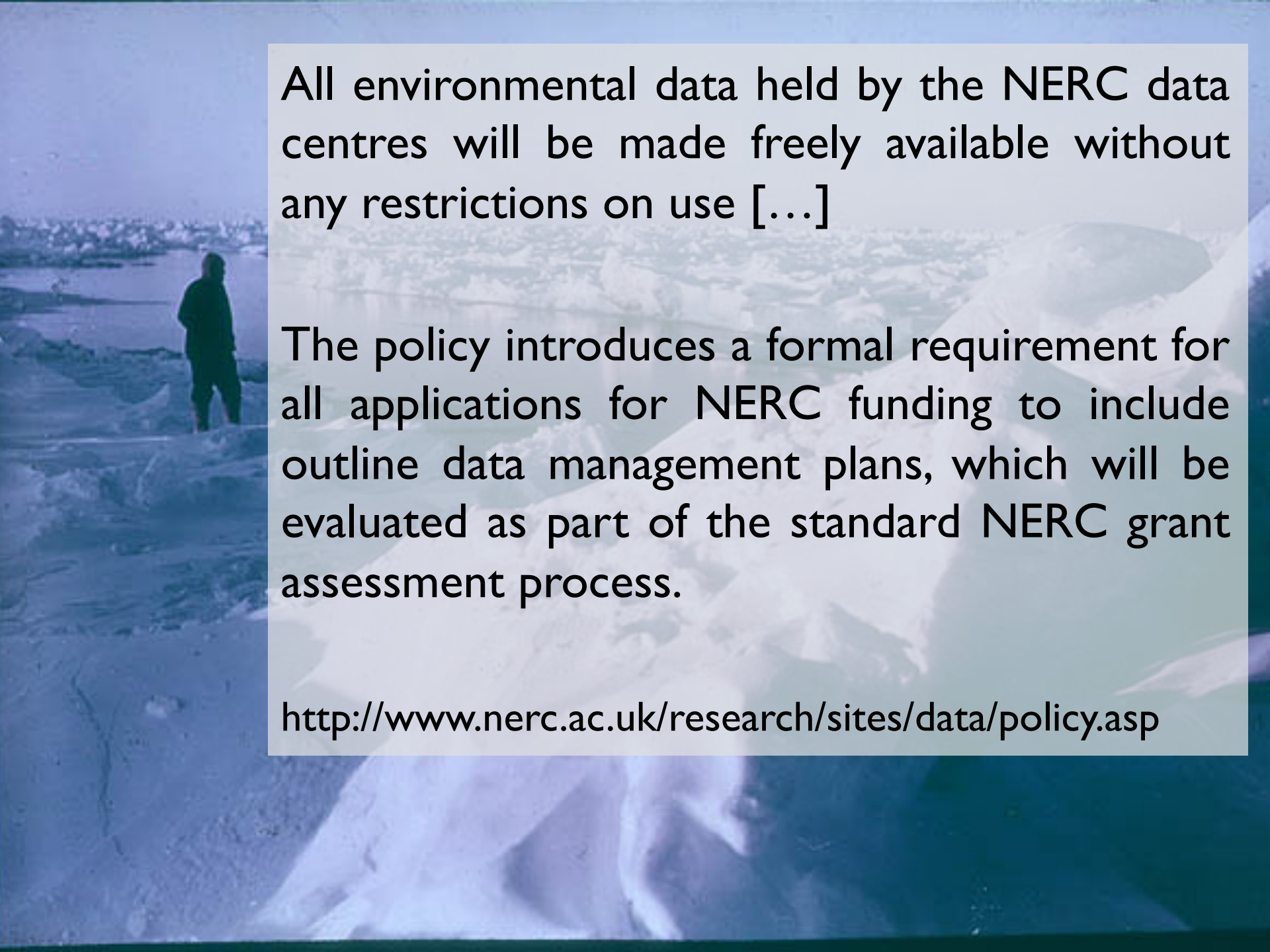


Odd one out ?



**NATURAL
ENVIRONMENT
RESEARCH COUNCIL**



A person in dark clothing stands on a vast, white, textured surface, likely snow or ice. In the background, a body of water is visible under a pale sky. The overall scene is cold and desolate.

All environmental data held by the NERC data centres will be made freely available without any restrictions on use [...]

The policy introduces a formal requirement for all applications for NERC funding to include outline data management plans, which will be evaluated as part of the standard NERC grant assessment process.

<http://www.nerc.ac.uk/research/sites/data/policy.asp>

RASMUSSEN REPORTS POLL

Did scientists falsify research to support their own theories on Global Warming?

59%

SOMEWHAT LIKELY

35%

VERY LIKELY

26%

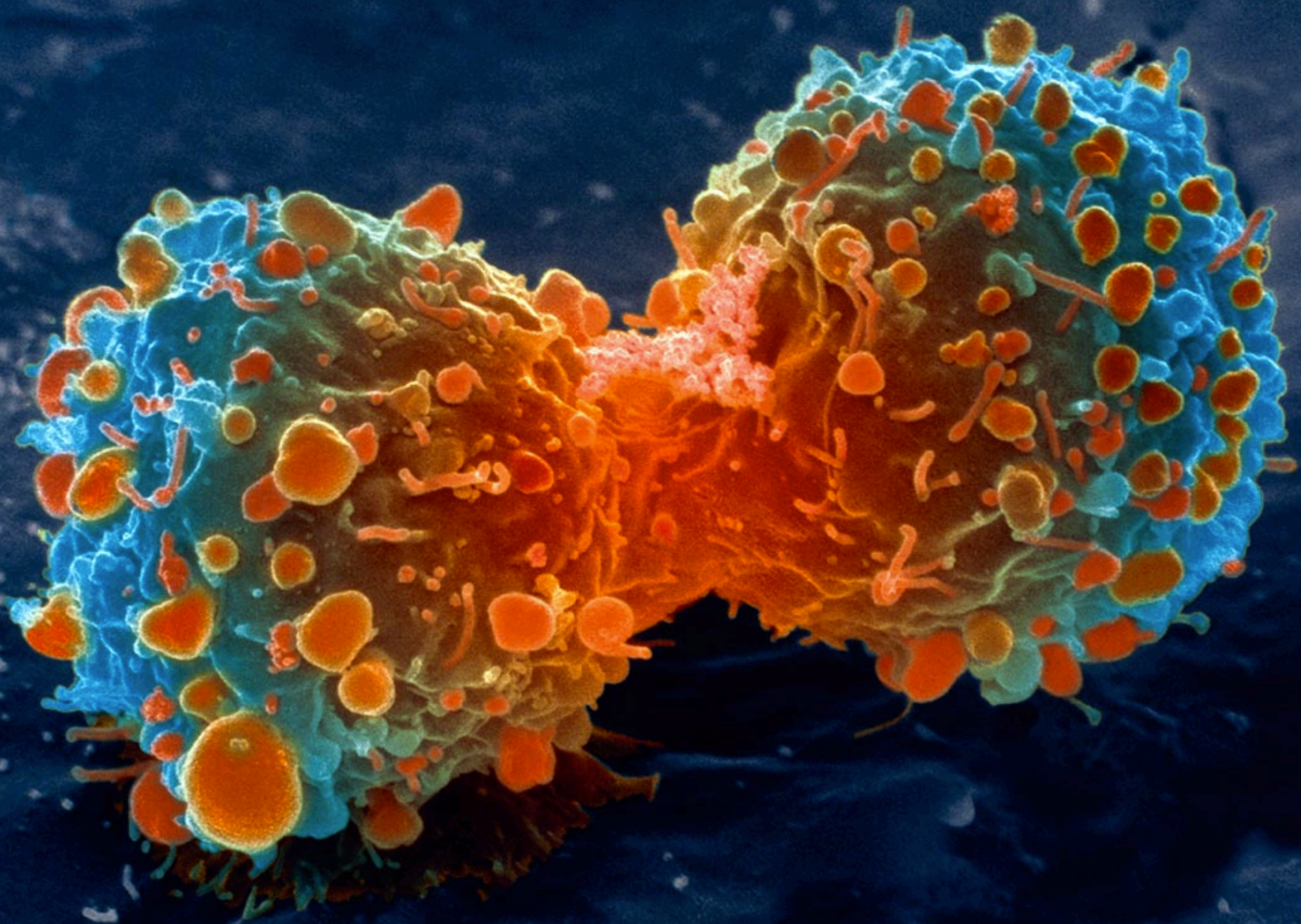
NOT VERY LIKELY




CLIMATE CHANGE RESEARCH / FOX NEWS \ GOP S NHL TOR 6 COB 3

Social accountability





A microscopic image of cells, likely from a tissue sample, showing various cell types and structures. The cells are stained, with some appearing in shades of orange and blue. A semi-transparent white text box is overlaid on the image, containing text about data sharing and NIH policies.

We believe that data sharing is essential for expedited translation of research results into knowledge, products, and procedures to improve human health.

The NIH endorses the sharing of final research data to serve these and other important scientific goals. The NIH expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers.

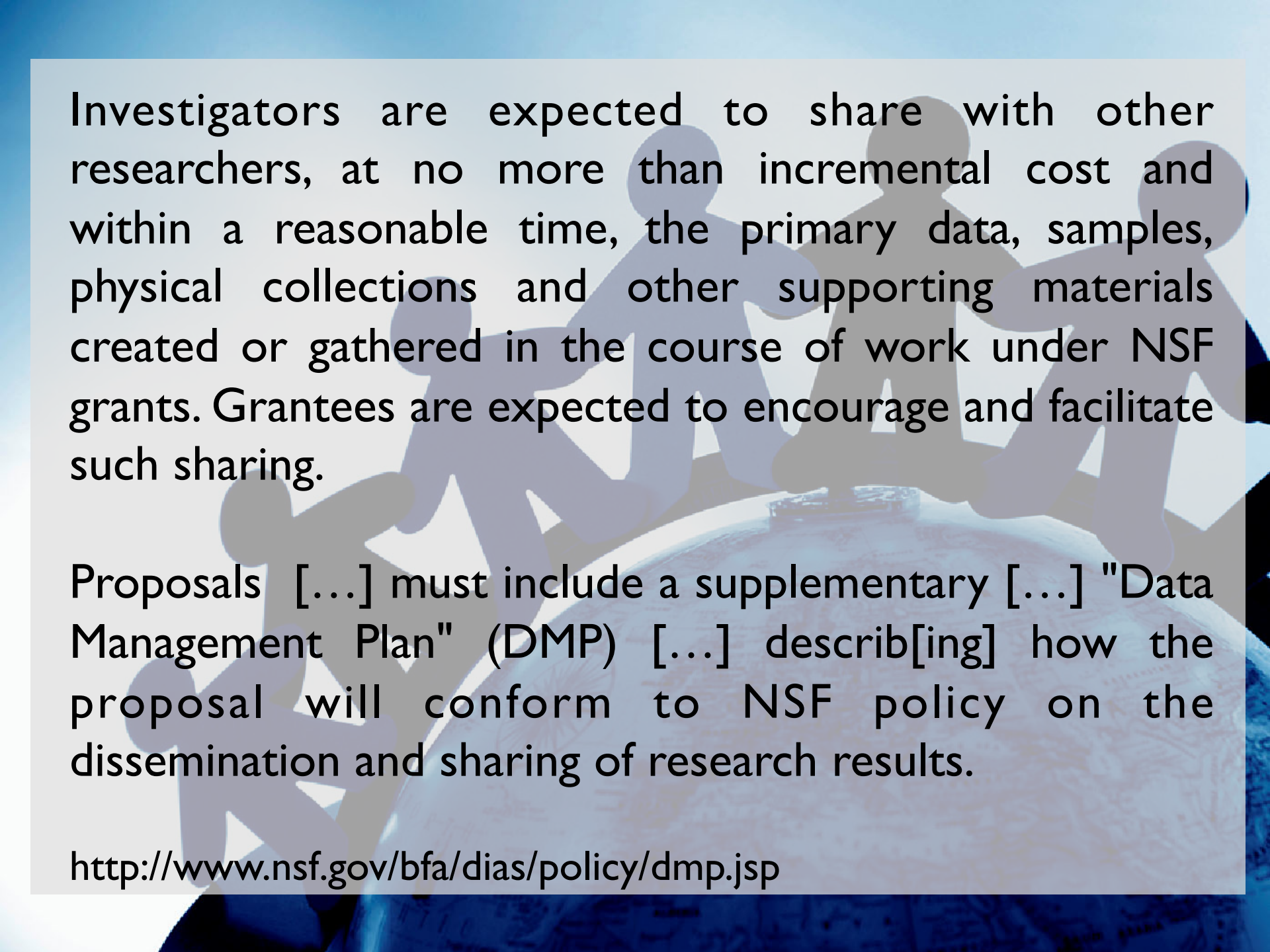
<http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html>



Social accountability







Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing.

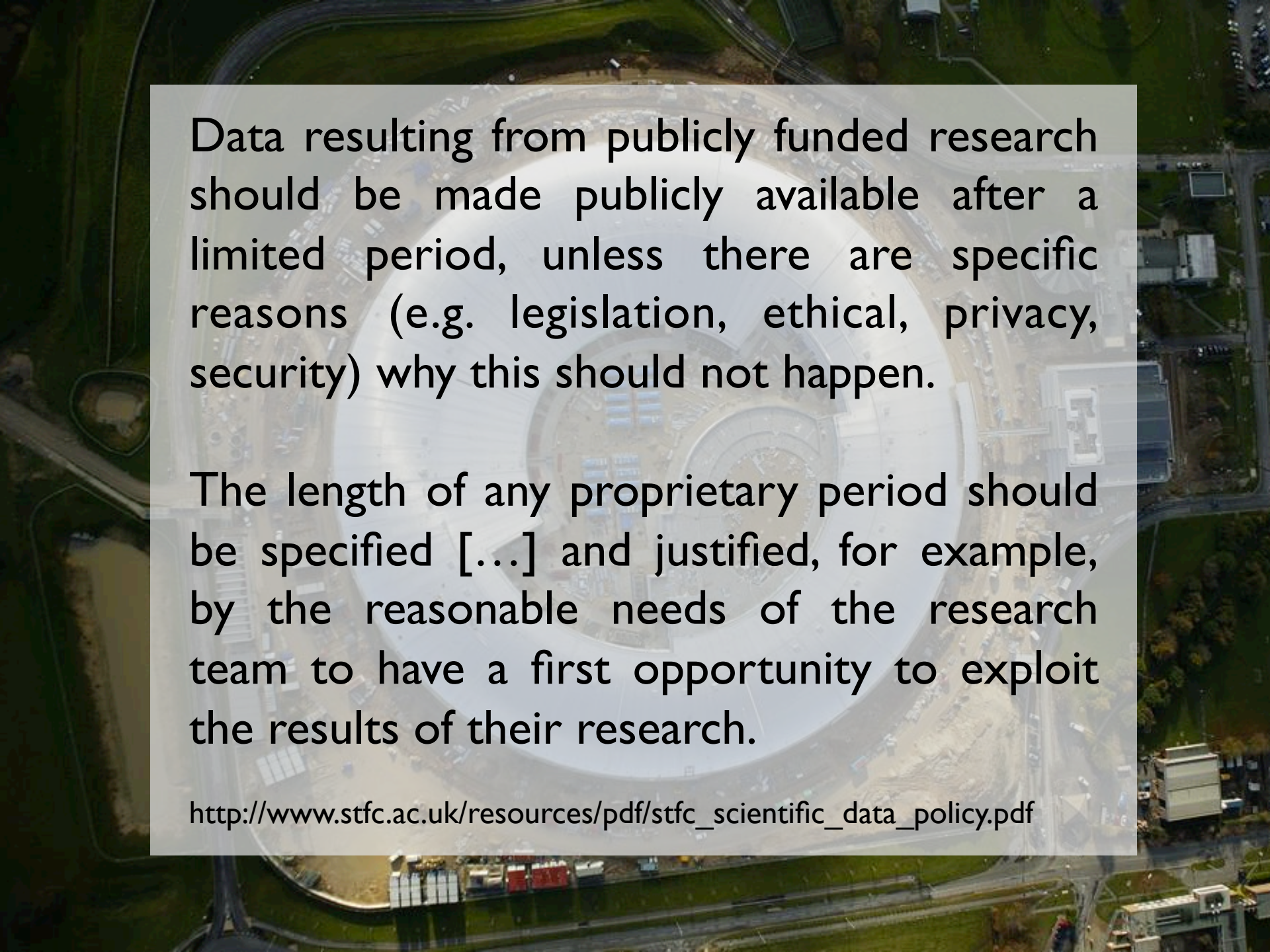
Proposals [...] must include a supplementary [...] "Data Management Plan" (DMP) [...] describ[ing] how the proposal will conform to NSF policy on the dissemination and sharing of research results.

<http://www.nsf.gov/bfa/dias/policy/dmp.jsp>



Science & Technology
Facilities Council





Data resulting from publicly funded research should be made publicly available after a limited period, unless there are specific reasons (e.g. legislation, ethical, privacy, security) why this should not happen.

The length of any proprietary period should be specified [...] and justified, for example, by the reasonable needs of the research team to have a first opportunity to exploit the results of their research.

http://www.stfc.ac.uk/resources/pdf/stfc_scientific_data_policy.pdf





**SLOW
MINEFIELD**





CERN Scientific Information Service supported CMS policy with knowledge built through FP7 EC projects: PARSE.Insight, ODE, APARSE

Convene informal group to “*harmonize principles for data preservation, re-use and (open) access at the LHC*”

Team member from the four collaborations’ DPHEP teams plus others suggested by each management



Marco Cattaneo (LHCb) Peter Clarke (LHCb)
David Colling (CMS) Kyle Cranmer (ATLAS)
Sünje Dallmeier-Tiessen (SIS) Tony Doyle (ATLAS)
Ulrik Egede (LHCb) Timothy Gershon (LHCb)
Mihaela Gheata (ALICE) Mike Hildreth (CMS)
Roger Jones (ATLAS) Kati Lassila-Perini (CMS)
Salvatore Mele (SIS) Lawrence Pinsky (ALICE)
Thomas Ruf (LHCb) Yves Schutz (ALICE)
Elizabeth Sexton-Kennedy (CMS) Lucas Taylor (CMS)
Roberto Tenchini (CMS) Mike Tuts (ATLAS)
Hans Von Der Schmitt (ATLAS)



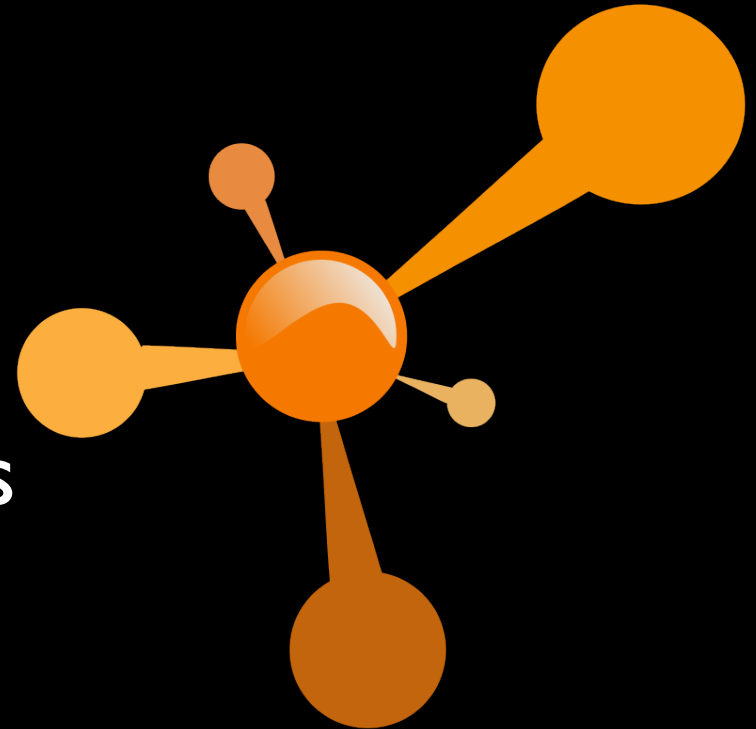
Consensus
benefits of harmonized principles



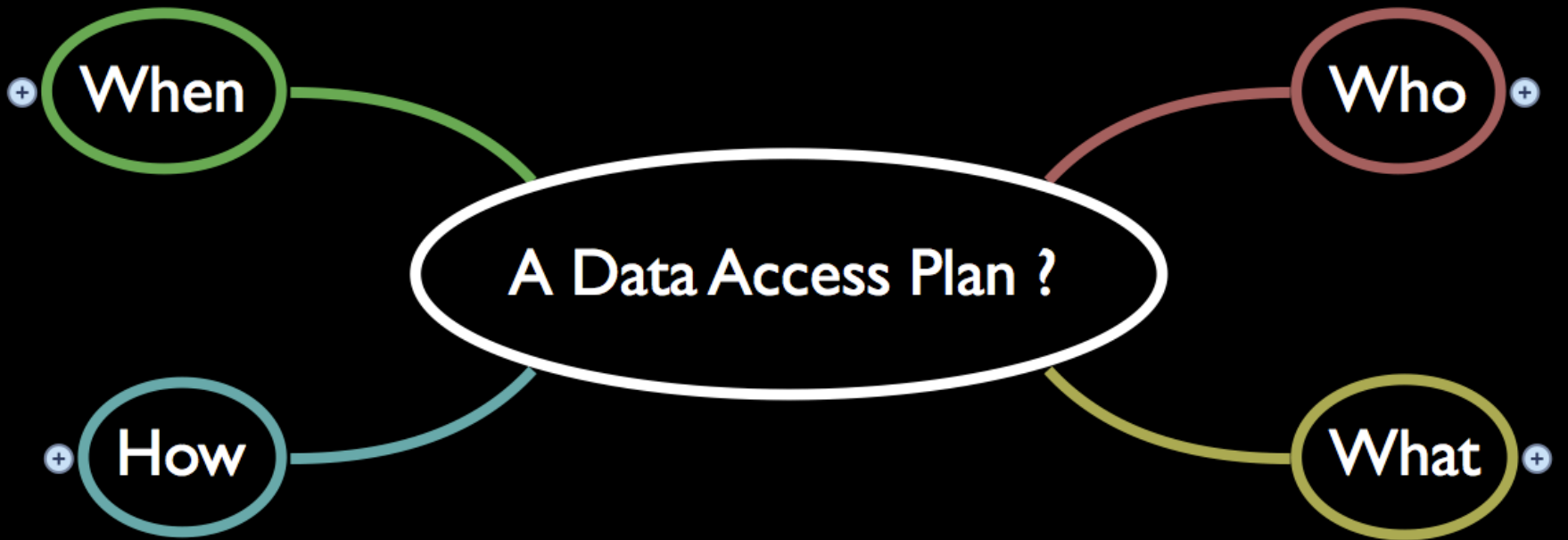
Funding agencies

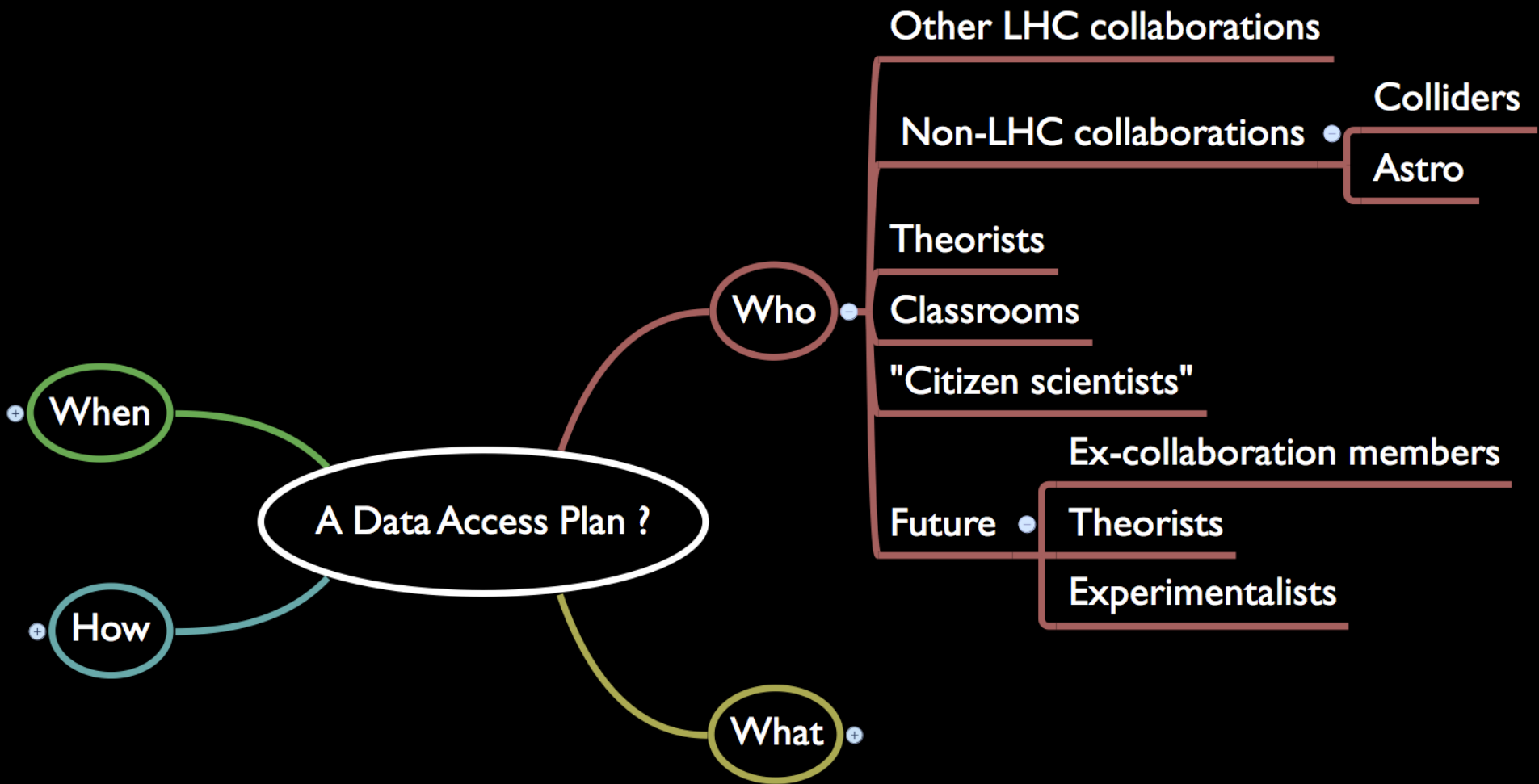
One experiment - Several funding agencies
One funding agencies - Several experiments

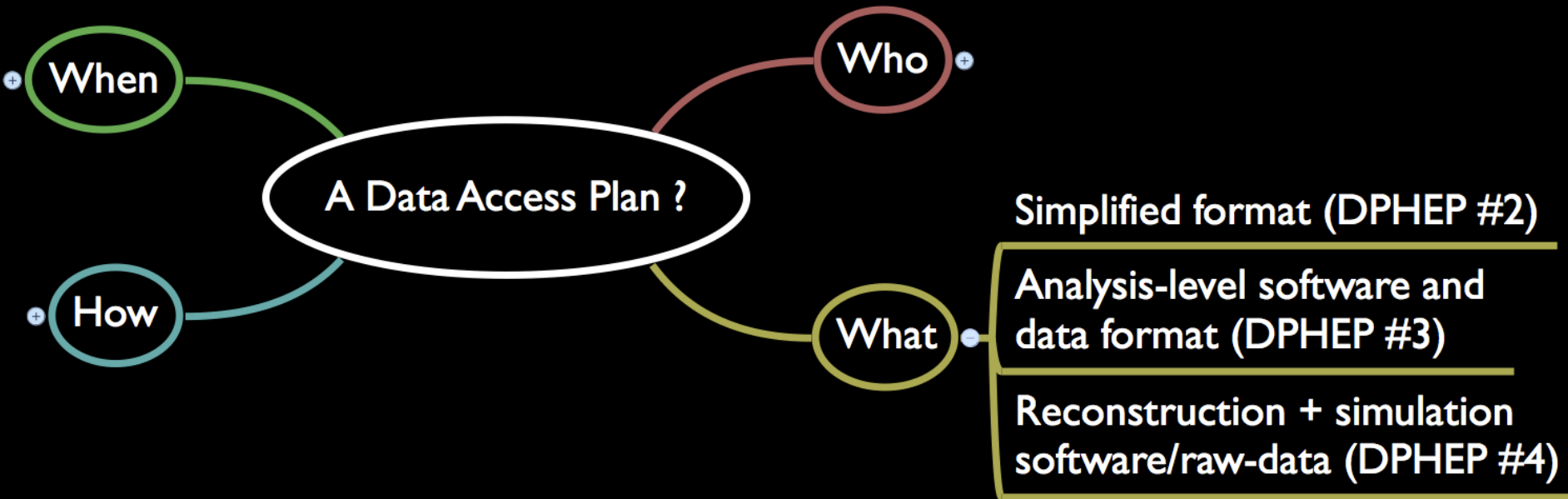
- Reaction to requests
- Pro-actively draft policies
- Coherent funding requests



Examples of discussion elements
(in each collaboration, and across the team)







Collaboration members

- Ad-hoc agreements
- "Open Access"

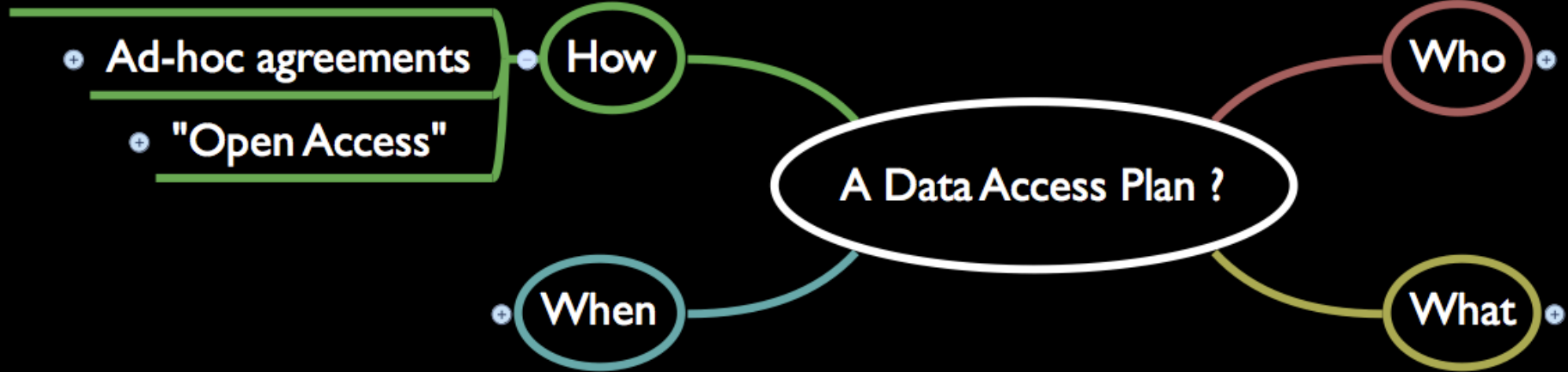
How

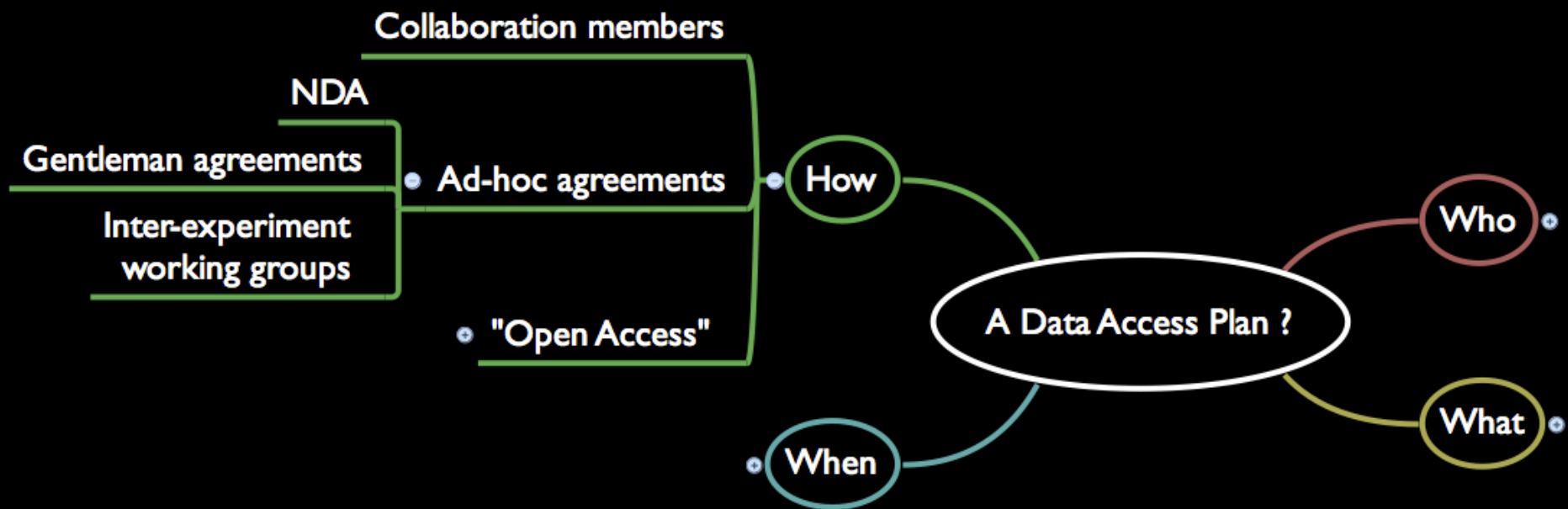
When

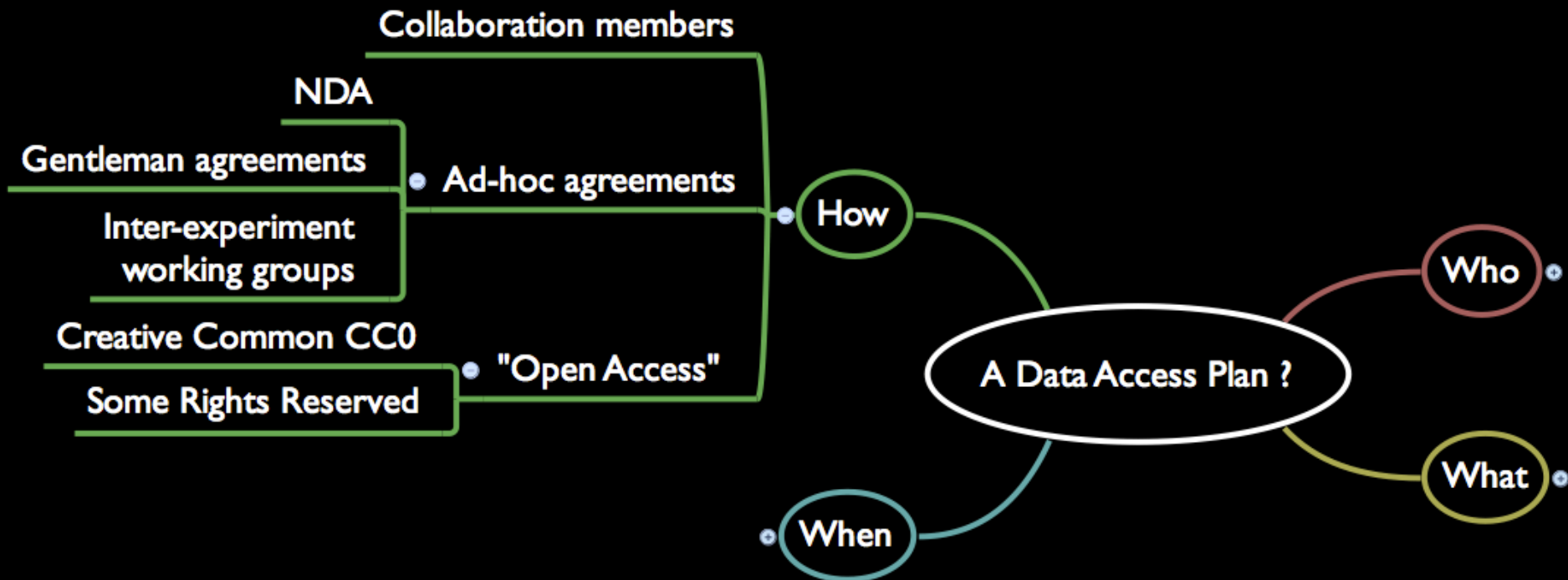
A Data Access Plan ?

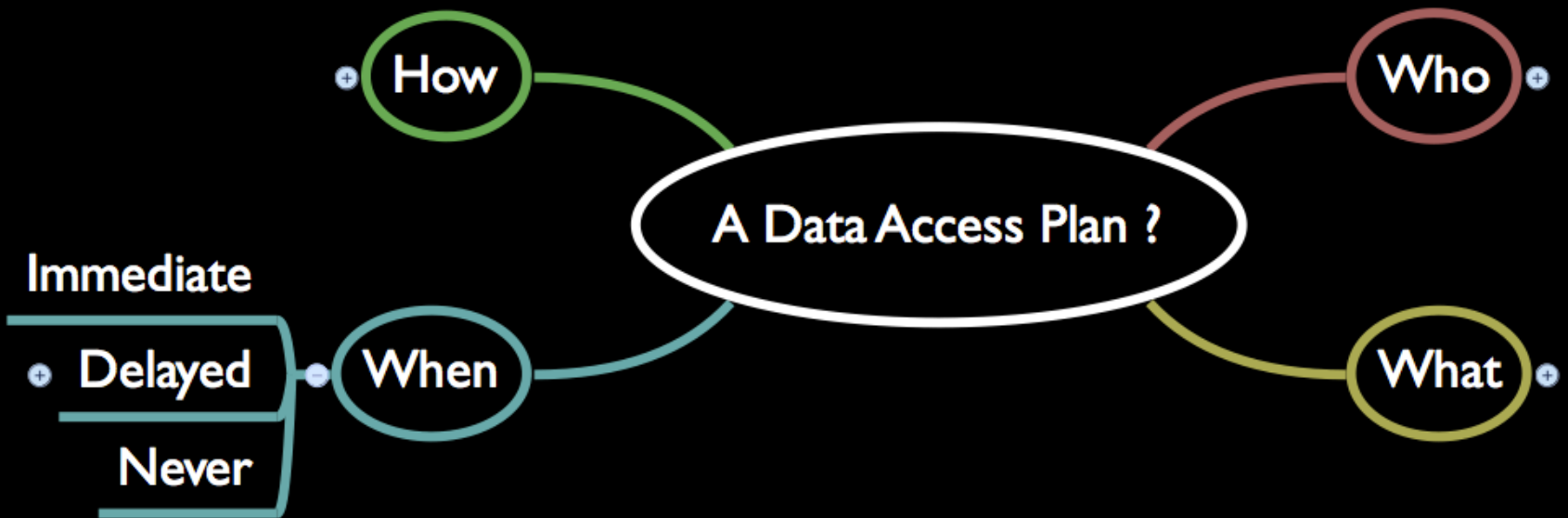
Who

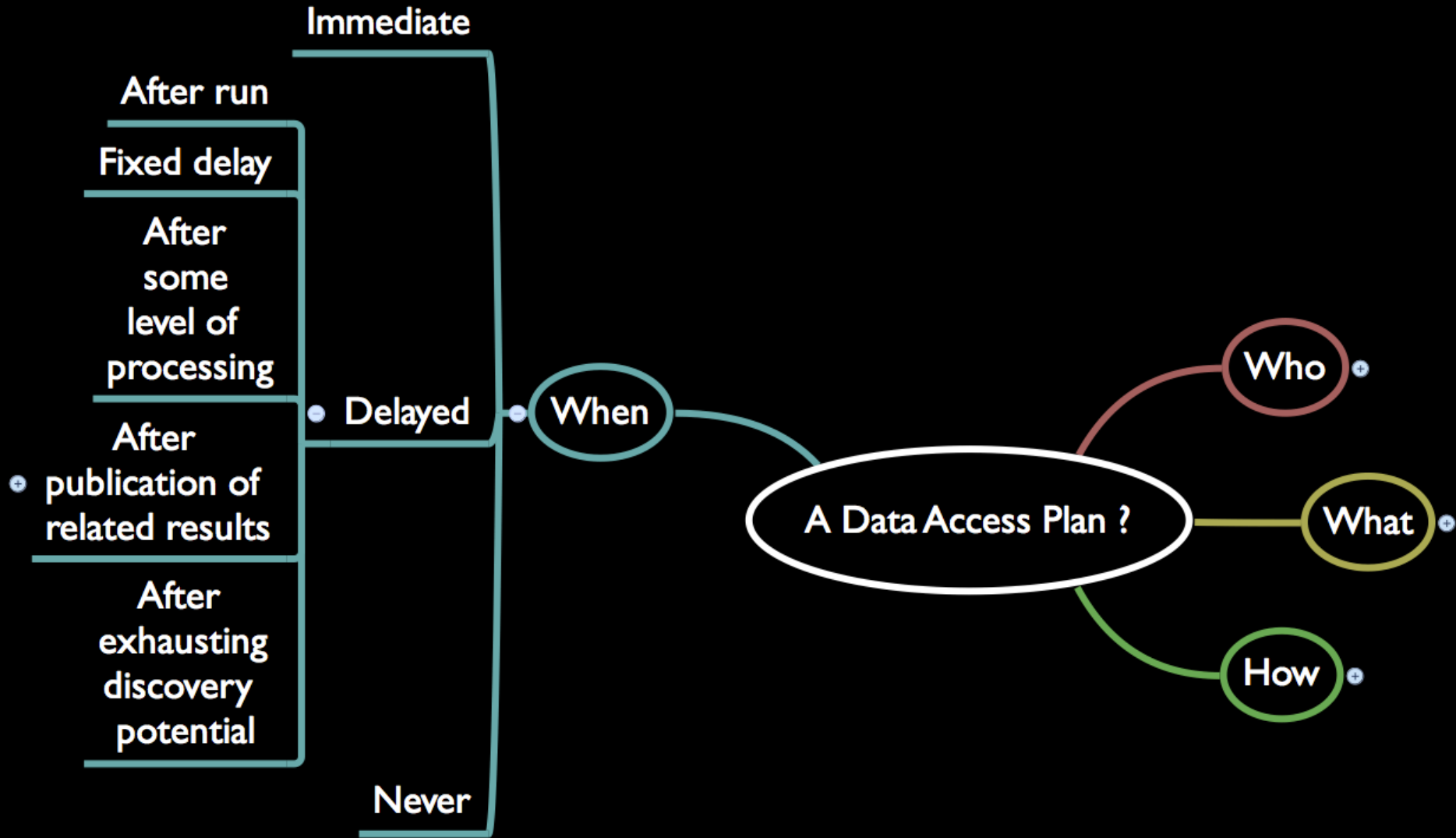
What

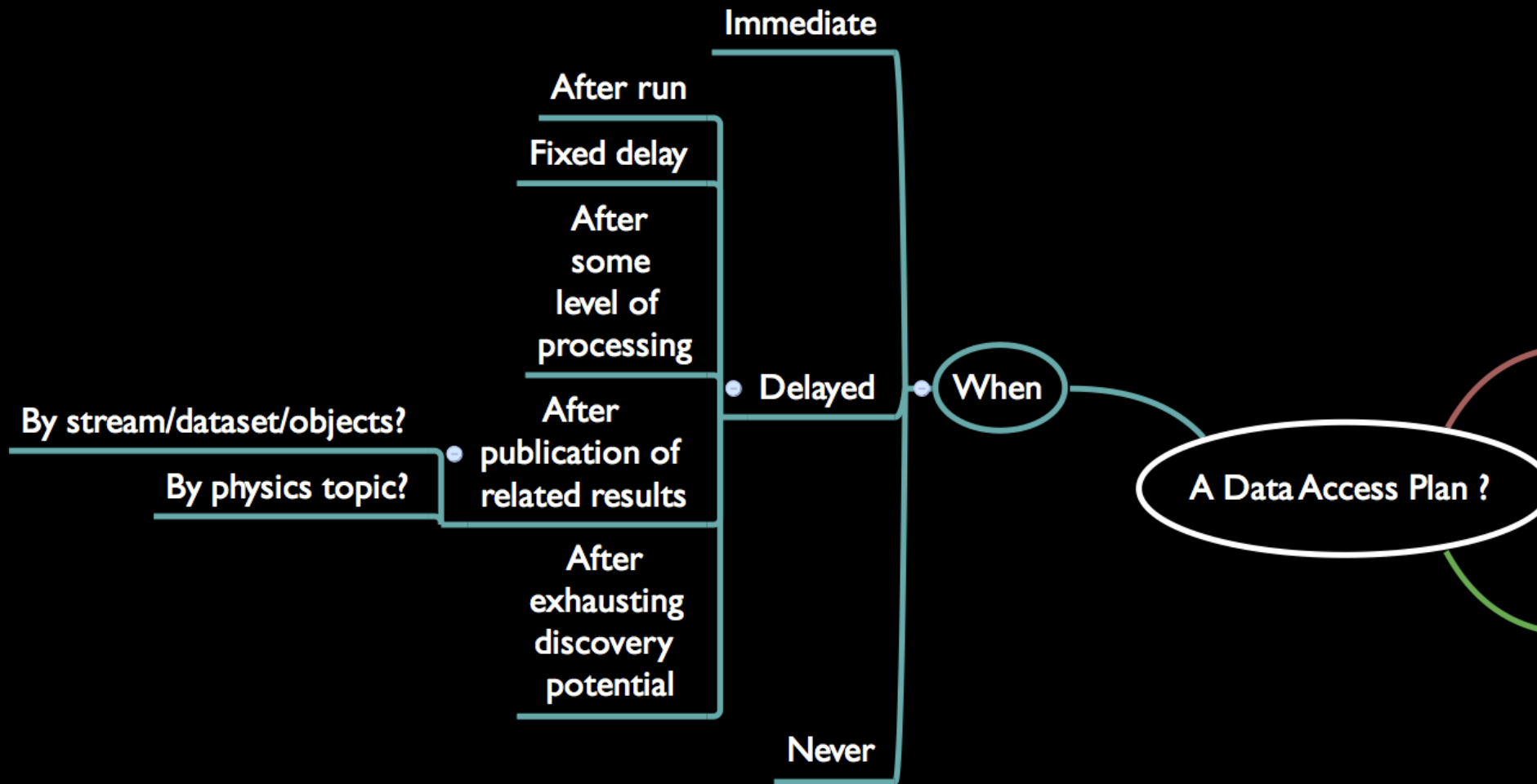












Consensus on eight principles

Capture the *status quo* of conversations within collaborations
who've written or are writing or are thinking
individual policies

1. The LHC Collaborations will assess which data, formats and software should be preserved to maximize the scientific potential of the LHC program, and prepare an appropriate policy. They will take the necessary actions to preserve that data and those formats, and the required environment to use them beyond the lifetime of the experiment. Such actions will require additional resources and external support.

II. The LHC Collaborations recognize the general principle that LHC data should become openly accessible at an appropriate time, both to scientists that did not participate to the experiments, and to any other interested party.

Those public releases, will use relevant standards for correct attribution and citation of data (DataCite DOI) and will adopt clear re-use license (Creative Commons CC0 Waiver).

III. Experts from the LHC Collaborations stress that the amount and timing of any open data release will depend upon the target audience, the level of complexity of the data and on the available resources, and there will be a compromise between the costs and the benefits. Such access need only be after those gathering and processing the data have had the opportunity to obtain the first benefit from it, as outlined in their project proposals. The LHC Collaborations should formalize these concepts in their policies.

IV. In addition to pursuing a full Open-Access policy to their publication, the LHC Collaborations will continue their policy of making openly available supplementary material (e.g. additional tables, figures, multi-dimensional distributions, etc) at, or close to, the time of publication. Existing infrastructures as HEPData and INSPIRE will be leveraged for this purpose.

V. The LHC Collaborations will continue activities to make extensive data sets at the four-vector level openly available for outreach purposes and conditional to dedicated additional resources.

VI. The LHC Collaboration will make openly available some fraction of their data in a detailed format suitable for detailed physics re-analysis, after some suitable embargo period, or else enable such analysis by other means. The amount of data and the length of the embargo will be individually determined by each Collaboration, depending, among other things, on the performance of the LHC accelerator. This will require resources above the current baseline experiment funding.

VII. Experts from the LHC Collaborations have identified raw data as having little practical use outside the collaboration, with principle VI above covering all practical re-use cases. Under exceptional circumstances, and within the purview of the Collaborations' governance, representative subsets of raw data may be made available in response to a well founded request.

VIII. Each Collaborations' governance structure will retain full responsibility for: setting and reviewing their data preservation, re-use and open access policies; implementing of an internal infrastructure to enact the policy; continuously monitoring and assessing the associated resource costs. Care will be paid to the interaction between the data availability policy and the collaboration policies on membership and authorship.

Next steps (?)