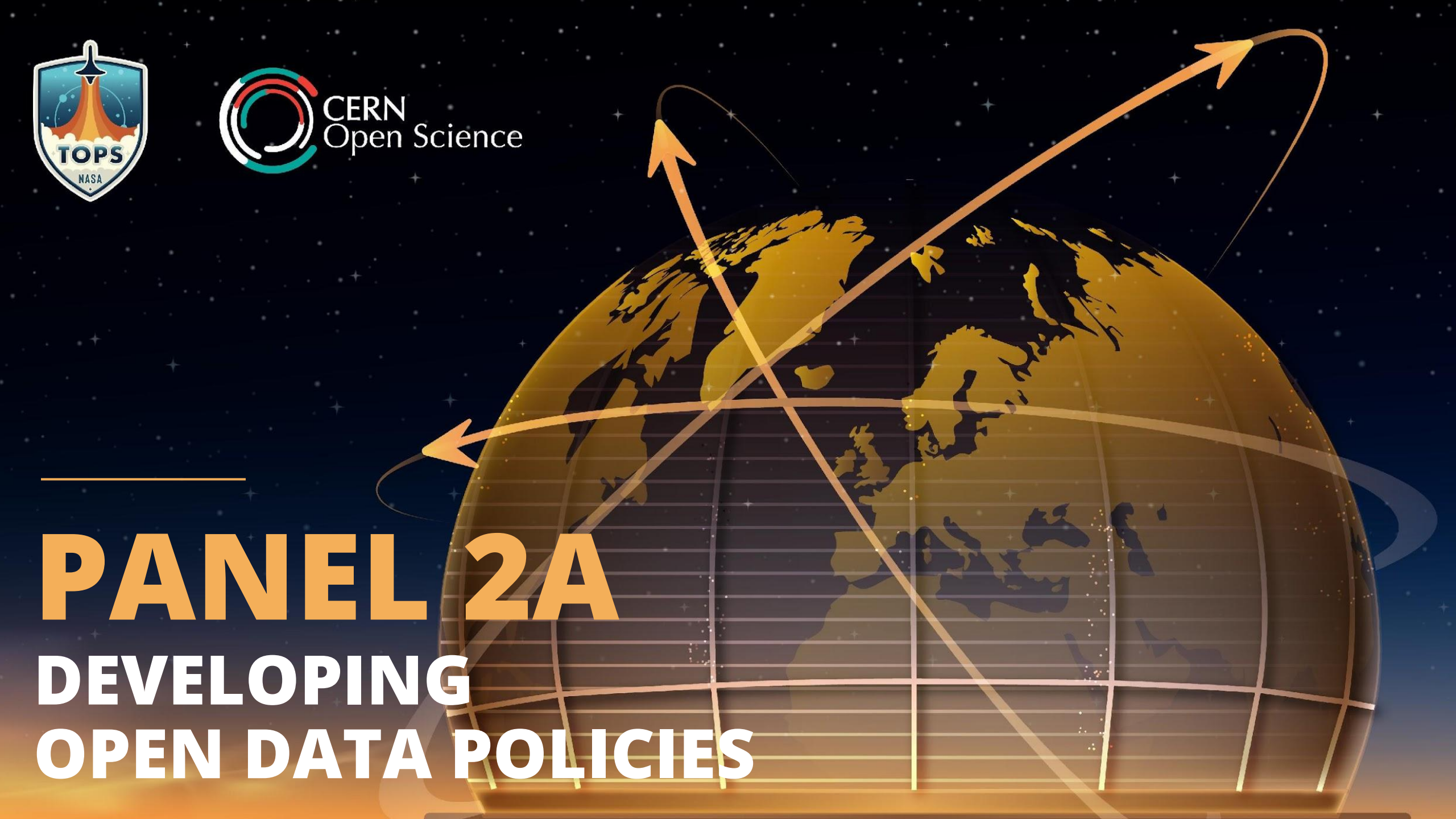


PANEL 2A

DEVELOPING OPEN DATA POLICIES

PANEL 2A

DEVELOPING OPEN DATA POLICIES

PANEL CHAIRS

Chris Marcum
Senior Statistician and Senior Science Policy Analyst, Office of the Chief Statistician of the United States

Nokuthula Mchunu
Deputy Director, The National Research Foundation of South Africa

PANEL SPEAKERS

Jamie Boyd
CERN Senior Scientist

Stefanie Lumnitz
EO Application Scientist, ESA

Rachel Paseka
Support Scientist, NASA Science Mission Directorate

The CERN LHC Open Data Policy

Jamie Boyd
Senior Scientist, CERN

Developed with:

ALICE

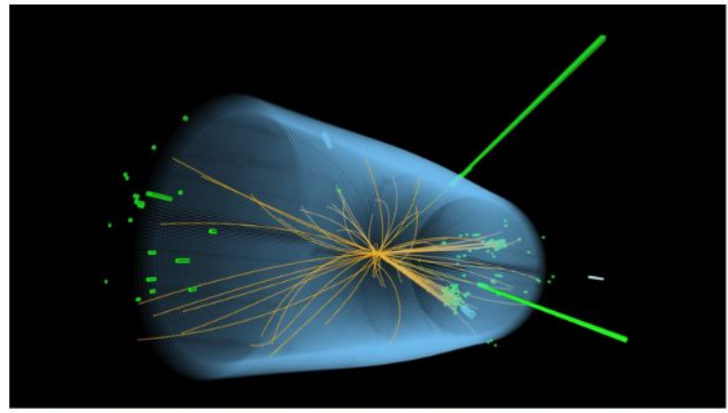
ATLAS EXPERIMENT

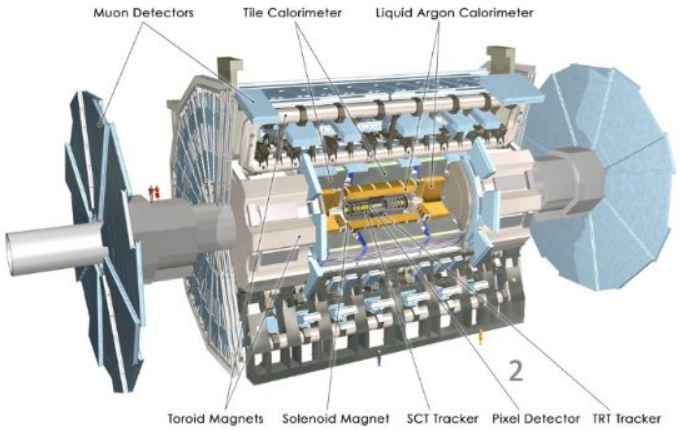
CMS

LHCb

Context

- LHC physics
  - Study high energy frontier (discovery of Higgs boson, search for new particles/forces etc…)
  - Complex detectors (millions of channels), large data volumes (O(50) PB/year raw data) etc..
- LHC community
  - 4 Large international collaborations (up to 3000 scientists / collaboration)
  - Collaboration lifetimes several decades
  - CERN as host laboratory
  - Collaborations have their own governance
- Increasing importance of Open Data
  - European Commission (relevant for funding applications)
  - European Stratgey for Particle Physics (2020 update)
- CERN management mandated working group to explore a common open data policy across the collaborations
  - To be formally endorsed by the Collaborations
- Expected community to use LHC Open Data
  - Professional physicists
  - Non-physics scientists (e.g. computer scientists: machine learning, reconstruction algorithms…)
  - Interested non-scientists





Challenges

- Concerns from the collaborations:
  - Ownership
    - Scientists who worked to design (R&D), build, operate the detectors unwilling to lose "ownership" of their data
  - Effort
    - Concern from experiment management that could lose effort to operate the experiment if people can analyze the open data without contributing to the experiment
      - Open Data policy of one LHC collaboration can effect other LHC collaborations => push for common policy!
  - Scientific rigour
    - Worry about lack of scientific rigour in analysis of open data (spurious claims)
  - Resources
    - Required resources within experiments (preparing open data datasets, documentation, storage space etc…), and from CERN side (person power and computing resources)

LHC data levels

- HEP Data Preservation (DPHEP) study group, divided particle physics data into 4 levels:

Small

- Level 1: scientific papers and associated auxiliary data
- Level 2: data tailored for outreach and education purposes
- Level 3: output of data reconstruction. The input for physics analysis.

Large

- Level 4: the raw data from the experiment

- Prior to the OD policy, all collaborations released Level 1, 2 data, and all agree that level 4 data is not useful for external bodies
  - Nearly all discussion was on Level 3 data

Main points in the policy

- Level 1 data:
  - Continue to release, including as much auxiliary data as possible to allow re-interpretation of the results (HepData database)
- Level 2 data:
  - Continue to release in appropriate formats/schedule
- Level 3 data:
  - Release data within 5 years after end of running period
    - Latency key to counter resistance from within the collaborations
    - Collaboration can withhold releasing data in special circumstances (unfinished high profile analysis ongoing)
  - Exact format determined by collaboration
  - Also release analysis software and simulated data samples
    - Needed to allow meaningful scientific study of the data
  - Documentation / support offered on best effort basis
  - Data released via CERNs OpenData portal
    - Storage media supplied by CERN (may not be long term solution, but for first 5 year period)
- Level 4 data:
  - Not useful, will not be released

CERN Open Data Policy for the LHC Experiments
November, 2020

The CERN Open Data Policy reflects values that have been enshrined in the CERN Convention for more than sixty years that were reaffirmed in the European Strategy for Particle Physics (2020)[1], and aims to empower the LHC experiments to adopt a consistent approach towards the openness and preservation of experimental data. Making data available responsibly (applying FAIR standards[2]), at different levels of abstraction and at different points in time, allows the maximum realisation of their scientific potential and the fulfillment of the collective moral and fiduciary responsibility to member states and the broader global scientific community. CERN understands that in order to optimise reuse opportunities, immediate and continued resources are needed. The level of support that CERN and the experiments will be able to provide to external users will depend on available resources.

This policy relates to the data collected by the LHC experiments, for the main physics programme of the LHC — high-energy proton–proton and heavy-ion collision data. The foreseen use cases of the Open Data include reinterpretation and reanalysis of physics results, education and outreach, data analysis for technical and algorithmic developments and physics research. The Open Data will be released through the CERN Open Data Portal which will be supported by CERN for the lifetime of the data. The data will be tailored to the different uses, and will be made available in formats defined by each experiment that afford a range of opportunities for long-term use, reuse and preservation. In general, four levels of complexity of HEP data have been identified by the Data Preservation and Long Term Analysis in High Energy Physics (DPHEP) Study Group[3], which serve varying audiences and imply a diversity of openness solutions and practices.

Published Results (Level 1) Policy: Peer-reviewed publications represent the primary scientific output from the experiments. In compliance with the CERN Open Access Policy, all such publications are available with Open Access, and so are available to the public. To maximise the scientific value of their publications, the experiments will make public additional information and data at the time of publication, stored in collaboration with portals such as HEPData,[4] with selection routines stored in specialised tools. The data made available may include simplified or full binned likelihoods, as well as unbinned likelihoods based on datasets of event-level observables extracted by the analyses. Reinterpretation of published results is also made possible through analysis preservation and direct collaboration with external researchers.

Outreach and Education (Level 2) Policy: For the purposes of education and outreach, dedicated subsets of data are used, selected and formatted to provide rich samples to maximise their educational impact, and to facilitate the easy use of the data. These data are released with a schedule and scope determined by each experiment. The data are provided in simplified, portable and self-contained formats suitable for educational and public understanding purposes; but are not intended nor adequate for the publication of scientific results. Lightweight environments to allow the easy exploration of these

CERN-OPEN-2020-013
20/11/2020

[1] European Strategy Group (2020), '2020 Update of the European Strategy for Particle Physics'.
[2] FAIR Guiding Principles for scientific data management and stewardship. Available at: https://www.go-fair.org/fair-principles/.
[3] Data management plans are defined by the LHC experiments to address the long-term preservation of internal data products. See: Akopov et al., Status report of the DPHEP Study Group: Towards a global effort for sustainable data preservation in high energy physics. arXiv preprint arXiv:1205.4667 (2012).
[4] Repository for publication-related High-Energy Physics data: http://www.hepdata.net.

1

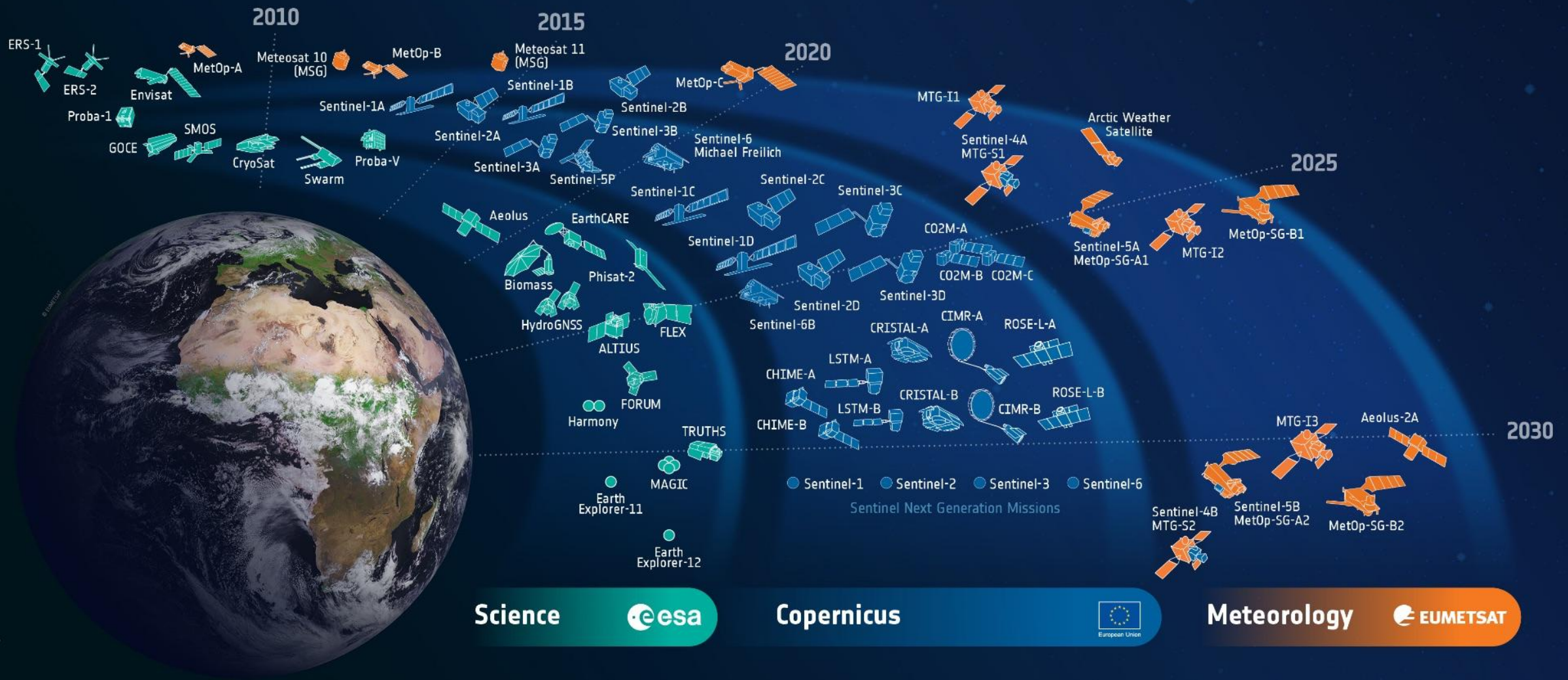
After ~1 year of discussions the policy was endorsed by the large LHC experiments in late 2020.
Work ongoing to encourage other CERN experiments to sign up to the policy

ESA EO Open Data Policy Landscape

Stefanie Lumnitz

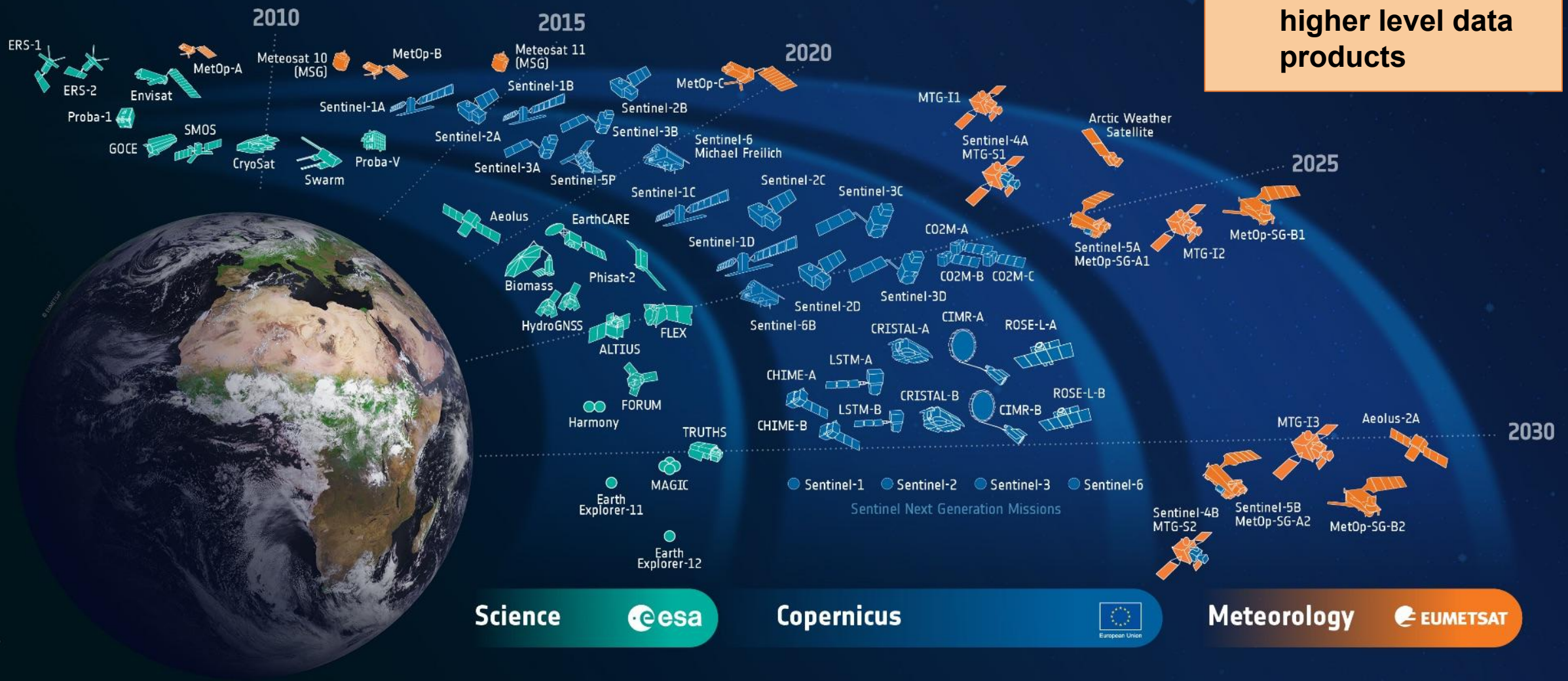
EO Application Scientist, ESA

ESA-DEVELOPED EARTH OBSERVATION MISSIONS



ESA-DEVELOPED EARTH OBSERVATION MISSIONS

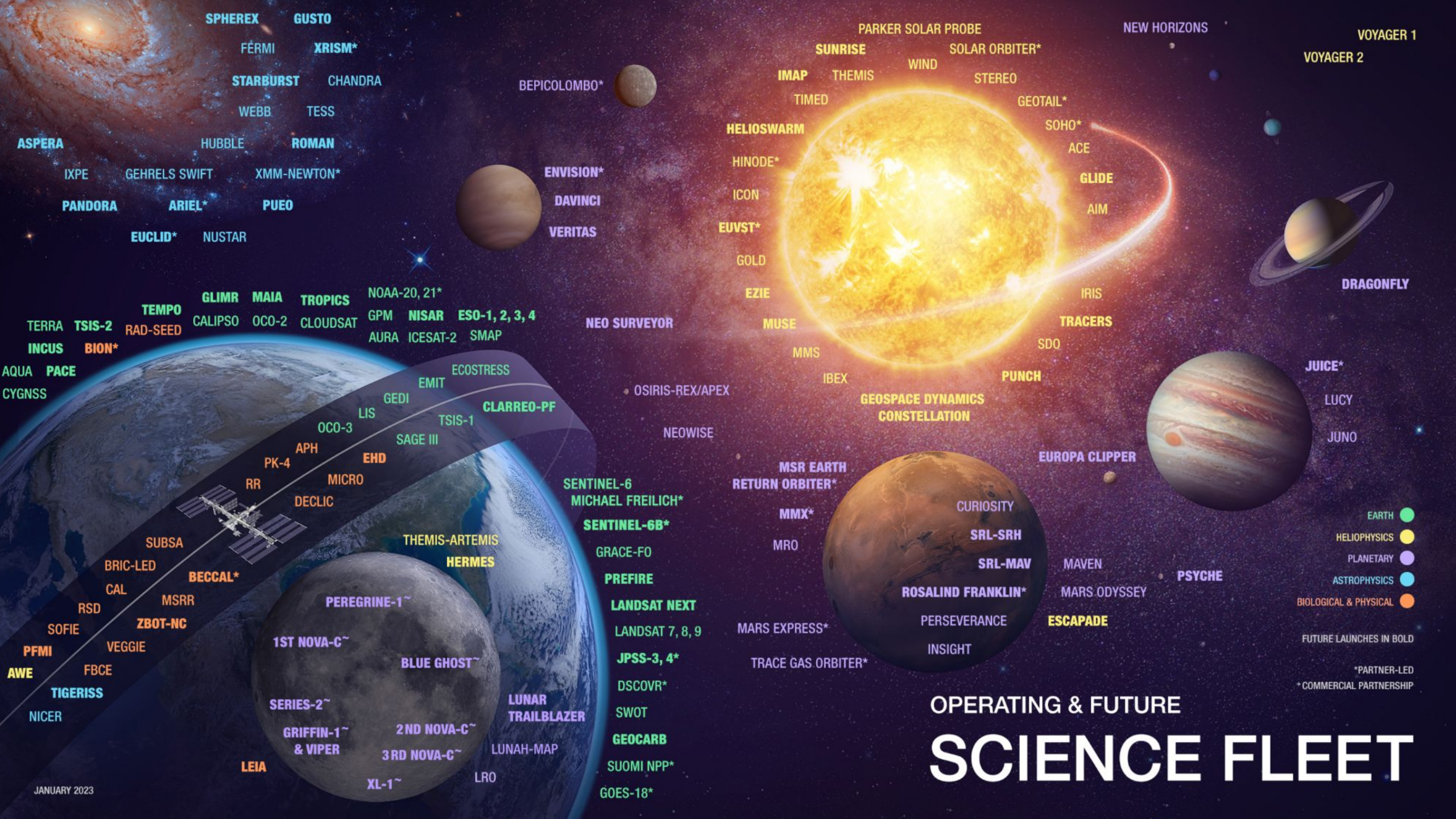
+ **Downstream and higher level data products**



NASA Open Data Policies

Rachel Paseka

Support Scientist, Open-Source Science Initiative
NASA Science Mission Directorate


SPHEREX
GUSTO
FERMI
XRISM*
STARBURST
CHANDRA
WEBB
TESS
ASPERA
HUBBLE
ROMAN
IXPE
GEHRELS SWIFT
XMM-NEWTON*
PANDORA
ARIEL*
PUEO
EUCLID*
NUSTAR
BEPICOLOMBO*
ENVISION*
DAVINCI
VERITAS
PARKER SOLAR PROBE
NEW HORIZONS
VOYAGER 1
VOYAGER 2
SUNRISE
SOLAR ORBITER*
WIND
IMAP
THEMIS
STEREO
TIMED
GEOTAIL*
HELIOSWARM
SOHO*
ACE
HINODE*
GLIDE
ICON
AIM
EUVST*
GOLD
EZIE
IRIS
MUSE
TRACERS
MMS
SDO
IBEX
PUNCH
GEOSPACE DYNAMICS CONSTELLATION
DRAGONFLY
JUICE*
LUCY
JUNO
EUROPA CLIPPER
TEMPO
GLIMR
MAIA
TROPICS
NOAA-20, 21*
TERRA
TSIS-2
RAD-SEED
CALIPSO
OCO-2
CLOUDSAT
GPM
NISAR
ESO-1, 2, 3, 4
AURA
ICESAT-2
SMAP
NEO SURVEYOR
INCUS
BION*
AQUA
PACE
CYGNSS
ECOSTRESS
EMIT
GEDI
LIS
CLARREO-PF
OCO-3
TSIS-1
SAGE III
APH
PK-4
EHD
RR
MICRO
DECLIC
OSIRIS-REX/APEX
NEOWISE
SENTINEL-6
MICHAEL FREILICH*
SENTINEL-6B*
GRACE-FO
PREFIRE
LANDSAT NEXT
LANDSAT 7, 8, 9
JPSS-3, 4*
DSCOVR*
SWOT
GEOCARB
SUOMI NPP*
GOES-18*
MSR EARTH RETURN ORBITER*
MMX*
MRO
CURIOSITY
SRL-SRH
SRL-MAV
ROSALIND FRANKLIN*
PERSEVERANCE
INSIGHT
MARS EXPRESS*
TRACE GAS ORBITER*
MAVEN
MARS ODYSSEY
ESCAPADE
PSYCHE
SUBSA
BRIC-LED
BECCAL*
CAL
MSRR
RSD
ZBOT-NC
SOFIE
PFMI
VEGGIE
AWE
FBCE
TIGERISS
NICER
THEMIS-ARTEMIS
HERMES
PEREGRINE-1~
1ST NOVA-C~
BLUE GHOST~
SERIES-2~
LUNAR TRAILBLAZER
GRIFFIN-1~ & VIPER
2ND NOVA-C~
3RD NOVA-C~
LUNAH-MAP
LEIA
XL-1~
LRO
EARTH
HELIOPHYSICS
PLANETARY
ASTROPHYSICS
BIOLOGICAL & PHYSICAL
FUTURE LAUNCHES IN BOLD
*PARTNER-LED
+COMMERCIAL PARTNERSHIP
OPERATING & FUTURE
SCIENCE FLEET
JANUARY 2023

Open Data Policy Initiatives in 2023

- ➔ Scientific Information Policy for the Science Mission Directorate
- ➔ NASA Public Access Plan

SPD-41a: Scientific Information Policy for the Science Mission Directorate (SMD)

Released December 2, 2022

Scope

- SMD: Astrophysics | Biological & Physical Sciences | Earth Science | Heliophysics | Planetary Science
- All future SMD-funded scientific activities, including new missions and research awards

Context for Development

- Previous policy (SPD-41) released in 2021
  - Consolidated existing Federal and NASA policy on sharing scientific information
  - Incorporated recommendations from SMD Strategy for Data Management and Computing for Groundbreaking Science 2019-2024
- Policy updates in SPD-41a arose from:
  - Community input via workshops and public comments on draft policy
  - OSTP Memo on Ensuring Free, Immediate, and Equitable Access to Federally Funded Research
- Development led by the SMD Chief Science Data Office, with input from all SMD divisions through Open Source Science Initiative Council

Scientific Information Policy website

SPD-41a: Scientific Information Policy for the Science Mission Directorate (SMD)

Released December 2, 2022

Key Requirements for Scientific Data *(new requirements in italics)*

- Scientific data underlying peer-reviewed manuscripts shall be made publicly available no later than the publication of the peer-reviewed article.
- *Scientifically useful data associated with a research award shall be made publicly available no later than the end of the award*.
- Mission data shall be openly available with no period of exclusive access.
  - The period for data calibration and validation shall be as short as possible and shall not exceed six months.
- Scientific data *should follow FAIR principles,* shall be made reusable with a clear, open, and accessible data license*, and shall be citable with a persistent identifier.*
- All SMD-funded scientific activities shall include a data management plan.
- *SMD proposal reviews: peer reviewed data and software shall be recognized as having commensurate value as peer reviewed manuscripts*

Scientific Information Policy website

NASA Public Access Plan for Increasing Access to Results of Scientific Research

Public release: May 18, 2023

Scope

- Agency-wide: all NASA research, development, and technology programs

Context for Development

- Original NASA Public Access Plan released in 2014
- Updates arose from:
  - OSTP Memo on Ensuring Free, Immediate, and Equitable Access to Federally Funded Research
  - Major advancements in the adoption of open science practices since 2014
- Development led by NASA Office of the Chief Scientist, with input from all relevant directorates (e.g., Science Mission Directorate) and offices (e.g., Office of General Counsel)
- Feedback provided by OSTP was incorporated prior to public release

NASA Public Access Plan

NASA Public Access Plan for Increasing Access to Results of Scientific Research

Public release: May 18, 2023

Key Requirements for Scientific Data

- Scientific data underlying peer-reviewed manuscripts shall be made publicly available no later than the publication of the peer-reviewed article.
- All proposals or project plans submitted to NASA for scientific research funding shall include a Data Management Plan

Components of Plan Relevant to Open Data

- Updates to NASA Research Data Policies
  - e.g., NASA Policy Directive 2230.1 - Research Data and Publication Access
- Future work: Guidance and training, infrastructure, and compliance processes and metrics

Next Steps

- Plan is currently open for public comment and public webinar will be held July 17, 2023
- Feedback will inform implementation of the plan

NASA Public Access Plan

THANK YOU