ESGF Strategic Roadmap

Planning Document
Revised July 10, 2015

ESGF Executive Committee Board
Dean N. Williams (Chair, DOE)
Michael Lautenschlager (Co-Chair, DKRZ)
V. Balaji (Princeton University and NOAA/GFDL)
Luca Cinquini (DOE/NASA/NOAA)
Cecilia DeLuca (NOAA)
Sebastien Denvil (IPSL)
Daniel Duffy (NASA)
Ben Evans (NCI)
Robert Ferraro (NASA)
Martin Juckes (STFC)
Claire Trenham (NCI)

Executive Summary

This document describes the ESGF mission and an international integration strategy for data, database and computational architecture, and stable infrastructure highlighted by the ESGF Executive Committee. These highlights are key developments needed over the next five to seven years in response to large-scale national and international climate community projects that depend on ESGF for success. Quality assurance and baseline performance from laptop to high performance computing characterizes available and potential data streams and strategies. These are required for interactive data collections to remedy gaps in handling enormous international federated climate data archives. Appropriate cyber security ensures protection of data according to projects but still allows access and portability to different ESGF and individual groups and users. A timeline and plan for forecasting interoperable tools takes ESGF from a federated database archive to a robust virtual laboratory and concludes the document.

Primary focus and direction of the roadmap includes the need to:

- Update or refresh technologies implemented within the ESGF software stack such as adopting CoG, the new Web user front-end interface; completely re-writing the node manager, used for peer-to-peer node communications; improving the software installation, making it easier for administrators to selectively install components for their site; and updating security, user authentication and authorization for access to federated data and managed resources.
- Focus on network and data center resources by utilizing the resilient 100 Gbps science networks in the US and across the Atlantic to Europe for large scale data replication for the most requested data sets and server-side computing for data reduction, analysis and exploratory visualization.
- Support climate science projects that want their data federated throughout the community and with other projects' data for possible intercomparison and knowledge discovery. These are many and include heterogeneous data sets from simulation, observation, reanalysis, and many other related climate science endeavors.

Introduction

This document describes the strategic plan for the Earth System Grid Federation (ESGF), which is an international collaboration to create open-source software and infrastructure that powers the study of climate science. The mission of ESGF is to create and maintain a robust federated data grid for the international climate-research community with access to relevant data, information, analysis and visualization tools, hardware, and network capabilities to make sense of peta/exa-scale scientific data. ESGF facilitates advancements in climate science by providing:

1. An easy-to-use and secure federated web-based software data infrastructure for large climate data sets;
2. A flexible infrastructure that enables customization by participating data projects to address their specific requirements;
3. High-performance search, analysis, and visualization tools that ensure data accessibility for and usefulness to the climate research community;
4. Access to a broad set of data and tools for comparative and exploratory analysis; and
5. A virtual collaborative environment for diverse research and analysis tasks with large and varied data sets.

ESGF (http://esgf.llnl.gov/) is driven by a collection of independently funded projects that develop, deploy and maintain the necessary open-source software infrastructure to meet the above-mentioned goals. It is a successful international collaboration that manages the first-ever decentralized database for handling climate science data, with multiple petabytes of data at dozens of federated sites worldwide. ESGF's widespread adoption, federation capabilities, broad developer base, and focus on climate science data distinguish it from other collaborative knowledge systems. The ESGF distributed archive holds the premier collection of simulations, together with observations, and reanalysis data to support analysis of simulations. Making it the leading source for today's climate model data holdings—including the most important and largest data sets in the global climate simulation community. In the future, ESGF intends to widen its scope to include other climate related data sets such as downscaled model data, climate predictions from both operational and experimental systems, and other derived data sets.

The ESGF production environment supports multiple international climate projects, including the WCRP Coupled Model Intercomparison Project (CMIP), whose protocols enable the periodic assessments carried out by the Intergovernmental Panel on Climate Change (IPCC). The data holdings and services in ESGF are distributed across multiple sites (such as BADC, DKRZ, IPSL, LLNL, NASA, NCI, NOAA, the Asian communities, and many more).

ESGF has greatly amplified the value of numerical climate model outputs and climate observations for current and future climate-assessment reports. However, the ESGF team faces substantial technical challenges due to the rapidly increasing scale of climate simulation and observational data, which is expected to grow from tens to hundreds of petabytes in the next five years. In a world of exponential technological change and rapidly growing sophistication in climate data analysis, ESGF must constantly evolve to remain useful. Fortunately, ESGF's well-defined governance structure helps to ensure that ESGF is advancing in directions that are most relevant to its supported user communities.

For more information on the current state of ESGF, see the Annual Earth System Grid Federation and Ultrascale Visualization Climate Data Analysis Tools Conference Report (Lawrence Livermore National Laboratory, Livermore, CA, LLNL-TR-666753; available online: http://aims-group.github.io/pdf/2014-ESGF_UV-CDAT_Conference_Report.pdf).

Background and Mission

ESGF is a federated data infrastructure that focuses on data-intensive science applications. Special emphasis here is on climate science, though basic principles are applicable to other data-intensive scientific disciplines as well. ESGF's overall mission is to improve the scientific workflow in climate science, including:

- Fostering international scientific cooperation;
- Intensifying global scientific data exchange;
- Improving globally federated scientific workflows;
- Facilitating traceability of scientific results; and
- Preserving data producers' visibility in data-intensive science applications.

Improvement of current scientific workflows is necessitated by a rapidly changing data landscape. The volume of climate research data has increased to a size that one single data archive cannot fit most scientific workflow requirements. Sharing of data management responsibilities and resources across a global federation is the goal. Requirements for data access and processing velocity will become more demanding and require data federation-based solutions. Along with the increase of data volume and variety comes a strong demand for climate research data management. A huge number of different data types will probably not be a pressing problem, but the huge number of individual data entities in the federation will be.

In climate research, a volume increase over the next five years from tens to hundreds of petabytes is expected, while the expected number of individual data entities (files) increases from millions to hundreds or thousands of millions. Based on this projection, the speed for data access and processing may increase from hours and days to weeks and months. Based on the classical 5V definition (volume, velocity, variety, veracity, value) for Big Data, climate research has a Big Data problem now, and it will only get worse over the next five years.

This huge data volume brings with it substantial challenges to current science investigation paradigms. The data is becoming too large to move over current or anticipated networks. This is forcing the infrastructure to provide services to either allow users to select and retrieve only the specific data subsets of interest, or provide the analysis capability within the infrastructure so that users need not move data to their workstations. This second option is not a trivial instantiation of analysis capability at the data sources, since the data sources themselves are remotely distributed. Data movement—both from the data sources to the end user, and among data sources and analysis capabilities—is inevitable and affects the entire data ecosystem design (see **Figure 1**).

In the long-term, the ESGF global data ecosystem will move into a virtual laboratory. This involves the migration from a data-centric infrastructure into an information-driven infrastructure. It also affects the ESGF user community accessing data and information from climate modelers and climate model data analysts to determine climate impacts, adaptation, and vulnerability. In terms of the IPCC process, the ESGF focus is intended to move from the Working Group I focus to all three IPCC Working Groups. To achieve this, the

ESGF global infrastructure development roadmap will migrate from federated databases over a data ecosystem to a virtual laboratory for earth system science in its broader sense.

Initially, ESGF infrastructure requirements will be set by two user groups—data providers (climate modeling groups and associated data center) and data consumers (climate modelers and climate model data analysts). As an example, these two groups were traditionally members of the Working Group I. Though ESGF infrastructure implementation is constrained by technical limitations and available human resources, ESGF must evolve to deal with this changing landscape and climate research more broadly. This document lays out the strategic roadmap for near- and medium-term development in response to these realities (see **Figure 3**). In addition, the use cases for other working groups are significantly different than those of Working Group I. For this reason, ESGF must start expanding its architecture to accommodate those additional use cases that are typically more local to regional and higher spatial and temporal resolution.

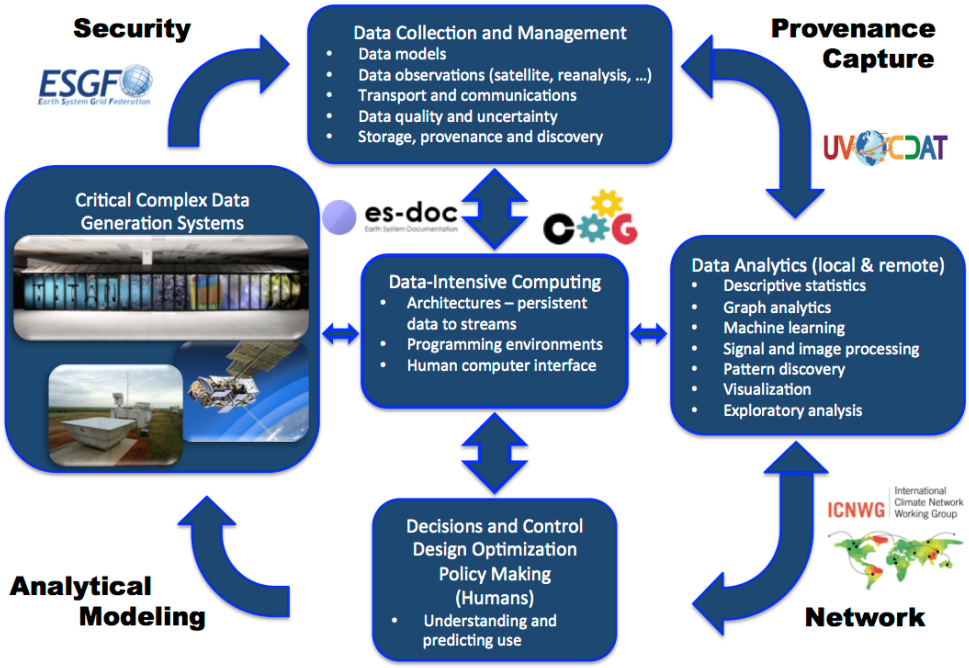

Figure 1. *The ESGF data ecosystem: The proposed components of the climate community's integrated data ecosystem and workflow currently under development, with comprehensive provenance capture. This integrated data ecosystem tightly interweaves every aspect of climate data research, from model development through interpretation and dissemination of research results via ESGF.*

Scientific Landscape and Overview

Earth's climate system science is about the understanding of all the components of the climate system (physical and biological) of the Earth. That is, how they function individually, and how they interact with and influence each other. Earth system research data are mainly about the physical domain and biogeochemical information:

- Output from numerical models;
- Observations from in situ networks, satellites, and aircrafts; and
- Analysis data products that merge observational data from multiple sources in real time.

Available data from these sources is easily in the tens of petabytes and will grow to hundreds of petabytes by 2020. Prior to the ESGF, researchers had a difficult time just identifying what data were available, much less how to access and interpret them. With the advent of the ESGF, certain classes of data (model output in particular) have become much easier to discover and retrieve. A large portion of that success is due to standardization on particular formats and conventions and the sharing of the infrastructure burden through common and consistent peer-to-peer services. This success has encouraged the broadening of the user base and data provider base, who now wish to expand the holdings of the ESGF, and the science that can be done with those holdings. ESGF data holdings are raw material for research in VIA communities (vulnerabilities,

impacts, and adaptation), and this broadens the scope for ESGF to scientists who are not familiar with climate model data handling.

The science drivers of this expansion are:

- *Scientific projects:* Data providers, who expect a reliable data infrastructure to make their products visible and accessible, while being able to control and track utilization and receive appropriate credit for their contributions. A main target of ESGF software infrastructure is to support WCRP internationally coordinated experiments and observations.
- *Scientific research teams:* Data users who expect easy discovery and access to data that is relevant to their investigation, including the information necessary to understand and use the data in an appropriate manner; this also includes the ability to share reproducible workflows and data results to the federation.
- *Model development and modelers:* Model development has among the most varied data management needs, which includes performing many small model runs with rapid turnaround during the model development phase, more computationally demanding uncertainty quantification and optimization work for model refinement, and massive data runs on leading supercomputers with the full array of ESGF features once the models are in production, like in internationally coordinated numerical experiments within WCRP. These models represent the coupled model (the integration of individual component models, such as atmosphere, land, and ocean and sea ice) or individual component models. In either case, the modelers expect to utilize shareable reproducible/repeatable workflows and data access from many different heterogeneous data sources for their model development.
- *Modeling and data centers:* Modeling and data centers must easily deploy and maintain the data ecosystem on high performance computing (HPC) hardware and networks for large-scale use—in particular, the tools, methods, and standards needed to support the community of projects and scientists.
- *Funding agencies:* Funding agencies would like to see improved scientific workflows and an adapted climate science data ecosystem that demonstrates cost reduction and demonstrable scientific benefits for their investments.
- *Application users and developers needing climate science data and information for insurance companies, water management, agriculture, energy, etc.*

Scientific Challenges

The use cases help to illustrate the challenges of model development and climate science and provide the requirements for an integrated data ecosystem. The first case illustrates the requirements for the intercomparison of many heterogeneous data sets. To quantify uncertainty in climate simulation results, scientists run ensembles of simulations and compare those ensembles with similar ensembles runs by other climate modeling centers. Large intercomparison projects, such as CMIP, form the basis for periodic assessments. In addition, simulation results for the present data and historical climate must be validated against relevant observational and reanalysis data. Dozens of intercomparison projects have formed and share many of the requirements outlined in the use case below. Model development and observations require remote access to data and tools. Different models use different grids and grid resolutions, thus requiring interpolated tools for in situ comparative analysis, diagnostics, and data access and archiving. These requirements will help to bridge the gap between model developers and individual climate assessment researchers.

Few scientists or groups have the resources to deal with petabytes of data, so most science investigations are built around subsets of multiple data sources—model data, observations, and analysis/reanalysis products. CMIP3 and CMIP5 provided the initial drivers for the ESGF development, and now the desire to compare models with observations is expanding the types of data that the ESGF is being tasked to handle. This is not only for model evaluation, but also for model diagnosis and development, driving a desire for higher frequency observations and corresponding model output. Regional models are evolving their own constituency, and taking advantage of the infrastructure for their own purposes, like for CORDEX. Data products representing Committee on Earth Observation Satellites (CEOS) ECVs (Essential Climate Variables) are making their way into the ESGF archives. Along with these new data types comes new user types, many of which are not experts in all the various data sets becoming available, and need the tools and the documentation to make appropriate and practical use of the wealth of data available.

The ESGF is also attracting non-traditional data providers—small groups and individuals with boutique products—that find the infrastructure an attractive and inexpensive way to serve their data to the community. The relative ease of becoming a data node on the ESGF has resulted in an expansion of data nodes from the original handful to over 40 today. This is making it easy to expose large quantities of climate science data to the research community but presents a challenge to the federation to maintain interoperable, secure and stable infrastructure, and quality data products. Users of the ESGF have come to expect that the data served is reputable and authoritative. Maintaining this expectation in the future is a major requirement. However, individual scientists would like to include their data into the federation for others to explore and discover, as it would increase their visibility and speed up model development and climate science research. For such cases, individual scientists' data sets will not meet the same rigor as those published by modeling or data centers. Thus, users of these data sets will use at their own risk—as the quality of the data indicated by the individual data provider.

The advancement of the data ecosystem must include modeling development, testing, and execution in its infrastructure for accelerating the state-of-the-science and bridging the gap between model development and climate research. The Accelerated Climate Model for Energy (ACME) project will be one of the first modeling projects, with a complete end-to-end workflow, to access ESGF's data archive, data analysis and visualization, workflow automation, and provenance capture features in expediting climate model development and research. The project will conduct simulations and modeling on DOE's most sophisticated HPC systems. In addition, it will focus on three key climate change science drivers involving atmosphere, land surface, and ocean and sea ice.

The climate science community has made large investments in existing ESGF tools. Integrating the needed use case capabilities above into the ESGF tool suite with familiar interfaces will further reduce the barriers for large-scale adoption. In addition, relatively simple interfaces are needed for other target audiences (adaptation researchers, students, etc.). With this type of environment, the broad community of researchers and modelers will be able to access popular data products in a highly transparent manner.

Stakeholders Interactions

Following the model established for the ESGF governance, ESGF stakeholders interact primarily with the ESGF Executive Committee, which then summarizes the feedback and reports to the ESGF Steering Committee (see **Figure 2**). Stakeholders can be grouped into four main entities/groups: ESGF as the data ecosystem, scientific projects as provider of raw data material, e-infrastructure projects providing a mash-up of services (including ESGF), and scientists as producers of scientific information. The four entities goals, needs, and duties include:

- ESGF: development, maintenance, and operation of data ecosystem
    - ESGF developers need featured roadmap and conformance to release management schedule. They benefit from well-defined procedures covering the design, development, and testing of their software, which ESGF Executive Committee (ESGF-XC) members and ESGF working team leaders must guarantee.
    - ESGF maintenance and operation shall be built on top of well-established procedures covering integration testing, upgrade procedures, and operations procedures. Maintenance and operations depends as well on the data management cycle applied by scientific projects, which ESGF-XC members, operational teams, and scientific projects must guarantee.
    - Priorities for work assignments will be set by the ESGF SC and enforced by the ESGF XC, as outlined in the ESGF governance document (http://esgf.llnl.gov/governance.html). These priorities will reflect the immediate needs of the projects sponsored/funded by the agencies.
- Scientific projects: definition, management, and operation of data generation projects
    - They provide qualified data that are ready to be ingested by ESGF. The definition phase must be done in close cooperation with ESGF-XC to ensure compatibility with ESGF software and data consistency across the federation.

- During the lifetime of the scientific project, regular exchange will continue between ESGF-XC and project representatives to assess the quality of the operation for this project. If any corrective actions are needed, the ESGF-XC will have to update the ESGF software developer roadmap to schedule necessary changes or evolution.

- E-infrastructure projects: use ESGF services and data to build third parties' systems
  - The European Earth Observation Programme, Copernicus, will deploy an extensive data service infrastructure. The ESGF services will feed into the system developed by Copernicus. It is very likely that both ESGF-XC and ESGF Steering Committee (ESGF-SC) will have regular exchanges to help shape and build level cooperation between projects.
  - Climate and environmental scientists benefit from ESGF services following organized use of ESGF resources provided by their institution. Institutions build additional software layers on top of ESGF to enhance synergies between ESGF and their local systems and resources (other sources of data and computing capabilities). ESGF-XC will organize regular feedback opportunities regarding those activities to help shape ESGF future developments.
- Scientific research teams: use raw data material for scientific questions and answers
  - Whether following organized access and use of ESGF provided by their institution (IPSL, NCAR, GFDL, and probably others) or acting as individual scientists needing access to ESGF materials, scientists need the proper documentation and support for:
    - The cooperation between ESGF as an infrastructure and ESGF as multi-project data archive;
    - The nature of the data they are working with; and
    - The nature of the system they are using (ESGF).
  - Regular surveys must be organized by ESGF-XC to ensure proper feedback.

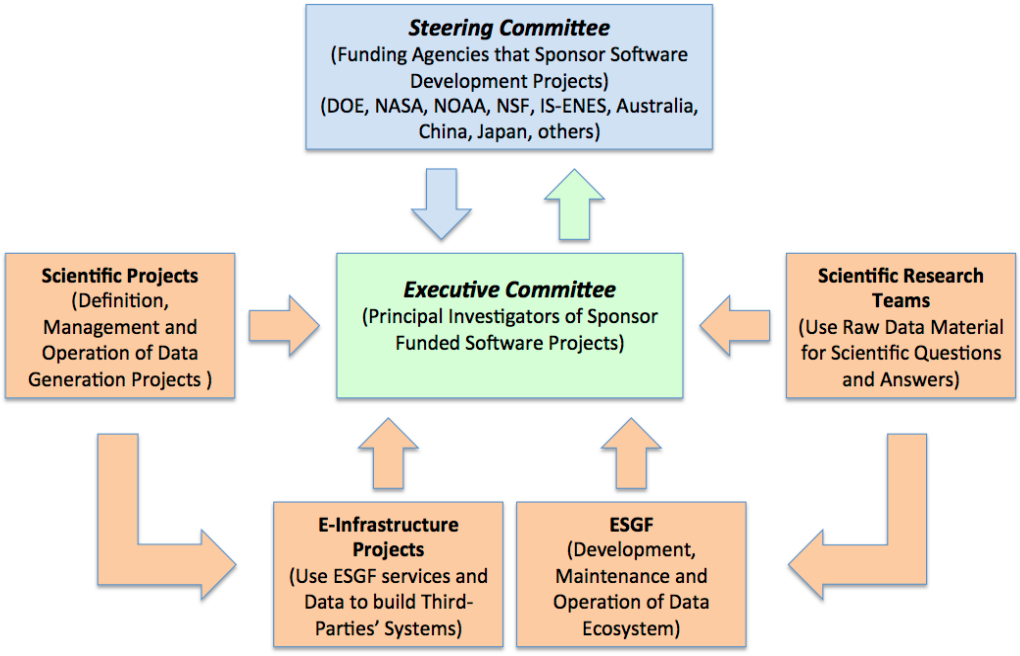

Figure 2. *Governance communication architecture: the different types of stakeholders (orange boxes) interact primarily with the ESGF Executive Committee (green box), which then reports to and receives guidance from the ESGF Steering Committee (blue box).*

Scientific Research Teams' Users' and Scientific Projects' Requirements for Climate Science Infrastructure

ESGF aims to develop and operate a software infrastructure for climate data that spans multiple centers around the world, stores data from different projects, serves a broad range of users, and delivers a wide range of services. It is also anticipated that medium to large projects will build downstream services and/or products partly based on ESGF services. Accordingly, the ESGF architecture and software stack must be designed according to some general principles that are intended to guarantee the best possible level of service to projects and users, as well as their long-term longevity and evolution. Specifically:

- *Federation of services:* A user must be able to search, discover, download, and analyze data hosted at different centers as if they were served from a single location. The distributed nature of the ESGF system must be totally transparent to end users and clients.
- *Unified access control:* In particular, a user or a software client must not be asked to authenticate or be authorized separately at all centers. Rather, the system infrastructure must support Single Sign On for authentication and federated access control, whereby the authorization statements issued by one center are honored by the other peer centers, for accessing the same class of resources.
- *Individual administration of local resources:* At the same time, a center or project must be able to define its own policies for accessing a certain class of resources (data, metadata, computing cycles, etc.). These policies must be propagated across the federation and consistently enforced.
- *No single point of failure:* The ESGF infrastructure must be designed so that interruption of services at one center will have minimal or no impact on the services offered by other centers. Resiliency, redundancy, and automatic failover must be built into the system both at the center and individual service levels.
- *Open-source software:* ESGF is a non-profit organization, funded with money from multiple government agencies that aims to support climate research for the benefit of all mankind. Therefore, all software developed by ESGF must be open source, a practice that also guarantees its highest possible quality, as software can be inspected, tested, and contributed to by the community at large.
- *Conformance to standards:* Whenever possible, ESGF software must conform to established standards for client-server and peer-to-peer interactions, in order to maximize interoperability with other agency systems and software packages. For example, some of the standards adopted by ESGF include OpenDAP, WPS, OpenID, SAML etc. Additionally, interoperability greatly increases the level of user satisfaction, as users are not compelled to learn and develop different techniques to access services from different systems.
- *Application Program Interfaces (APIs):* When existing standards are not available or not practical, ESGF services must be developed to conform to custom, well-documented APIs (for example, the ESGF Search API). This practice facilitates usage by users and clients, and allows for alternative implementations of the services to be developed or swapped entirely without any disruption for the users of the system.
- *Best software practices:* ESGF software developers must strive to apply recommended best practices in all phases of the software lifecycle (design, development, testing, deployment, operation) and across all software layers. This can be achieved by many collaborative events such as software code sprints, code reviews, and test coverage analysis.
- *Software modularity:* The ESGF software stack is not built as a monolithic package that must be installed and upgraded as a whole. Rather, it is based on the integration of several servers and libraries that are meant to be upgraded and possibly replaced individually. This philosophy enables the ESGF infrastructure to continuously evolve to incorporate new advances in all classes of services: data discovery, transfer, analysis, visualization, etc.
- *Scalability:* ESGF services must be designed to be able to scale to the order of magnitude of future data and metadata archives that are expected in the next 5–10 years, while still guaranteeing a satisfactory level of performance to the users. In particular, ESGF must be able to support the hundreds of petabyte-sized distributed archive that is expected to be generated by the next generation of climate models and higher resolution observing instruments.
- *Workflows and provenance:* The sheer size of current and expected future archives makes it impossible to store and analyze data on the users' personal workstations. ESGF must develop the capability of submitting complex data analysis workflows that seamlessly process data that are stored at distributed locations. The detailed workflow metadata (inputs, outputs, algorithms) must be

captured and made publicly available so that other researchers can fully understand and reproduce the results.

- *Networks for high-speed data transfers:* Specifically for CMIP6, hundreds of petabytes of data will need to be replicated among various data centers internationally. Data centers are expected to support high-speed data replication. GridFTP and its cloud-based derivative Globus are the suggested transfer protocol and service that will be implemented at the data center sites, since the existing HTTP-based tools used for ESGF will not scale to the data rates required for petabyte-scale data replication. In addition, ESGF continues to work with the international network communities to setup and tune Data Transfer Nodes (DTNs) in the Science DMZ[1] model (http://fasterdata.es.net/science-dmz/DTN/).
- *Proactive engagement with stakeholders:* The ESGF management and development teams must continuously and proactively engage with all possible classes of stakeholders—data users, data providers, project coordinators, infrastructure providers, and funding agencies. This will guarantee that the ESGF software is developed to fulfill the stakeholders' requirements, maximizes the users' satisfaction, and achieves the expected level of service.
- *Key performance indicators:* The ESGF software stack must include facilities for capturing and analyzing metrics about utilization of services, as well as for estimating the impact of the software infrastructure over the science community (for example, as quantified by the number of science papers that use some data sets downloaded from ESGF, or based on processing algorithms executed on ESGF servers). These metrics can be used to both improve the performance and quality of services and for reporting usage to the funding agencies.

It is recommended that projects and user groups form their own mechanisms for communicating requirements to ESGF—via the governance model shown in **Figure 2**. As an example, the World Climate Research Programme's Working Group on Coupled Modelling has appointed an infrastructure panel to articulate requirements for the CMIP data activities.

Roadmap and Timeline

The strategic development in ESGF focuses on the transition from the federated data infrastructure to an overall earth system science data ecosystem in order to eliminate projected Big Data challenges in the next few years and to ensure smooth operation of scientific workflows. The roadmap and timeline are shown in **Figure 3**. If approved by the ESGF governance, the implementation plan will layout the specific details of the roadmap.

[1] In networking, DMZs are used to create a perimeter network, which can enhance the performance and security of some applications. The acronym DMZ is a metaphorical meaning for "demilitarized zone."

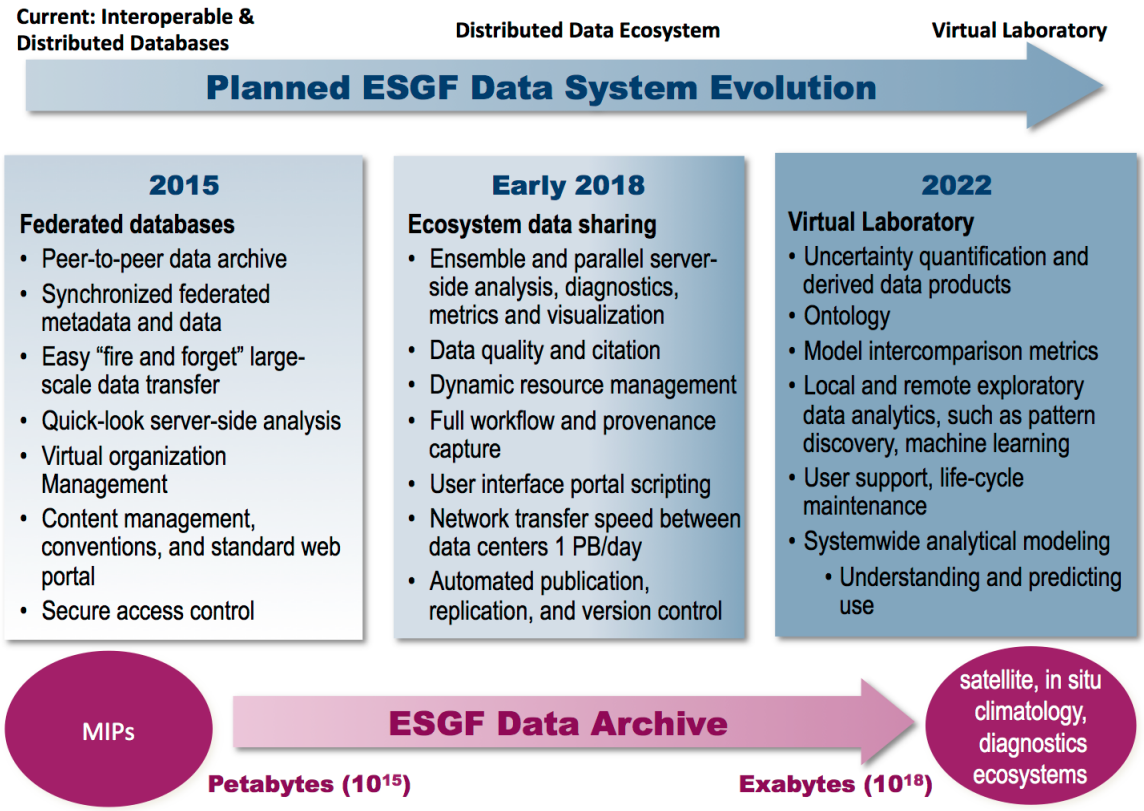

Figure 3. *ESGF data ecosystem roadmap and timeline migration from the current data federation to a virtual laboratory: a close collaboration between data scientists, climate scientists, and scientists from other domains is enabling the development of a data ecosystem that can support a broad variety of data and disciplines. Ecosystem features and fault tolerance will be incorporated into the ESGF system incrementally, over the next five to seven years.*

ESGF delivers a comprehensive, end-to-end, and top-to-bottom environment for current petascale and emerging exascale science domains. The figure emphasizes data services at each level for the node architecture. The production of ESGF for climate products is evidence that a distributed dynamic federated system is flexible enough to support a wide range of heterogeneous data (i.e., simulation, observation, and reanalysis) and application tools.

The long-term strategic development in ESGF reflects the necessary change of paradigm in climate research from data-centric to information-driven infrastructures. In the future, ESGF users will interact with data and other researchers through virtual laboratories that enable the natural evolution of climate based questions to be incorporated within the ESGF system itself. In other words, users will not just ask for climate data, but will be able to pose powerful climate based queries from understanding climate effects on weather phenomena (such as the future changes in hurricane statistics or regional changes of tropical days) or to predict future climatologies across the globe. In addition to elevating the level of research questions from data to information based queries, users will have a virtual environment in which to explore, manipulate, and visualize scientific information across ESGF. The provenance of these queries will be managed in a consistent manner with the data itself, enabling the publication, not of data sets, but of climate information.