Science-driven e-Infrastructure Innovation (SEI) for the Enhancement of Transnational and Interdisciplinary Data Use in Environmental Change Research

(Annex to Concept Note)

ANNEX 1: Examples of Relevant Problems

A set of examples[1] is provided here solely to illustrate the kinds of issues and barriers that might be addressed by this new Collaborative Research Action (CRA) funding Call. They are intentionally generic and cover some of the immediate challenges of tomorrow's transnational data-driven research that must be coupled, interdisciplinary and lead to real, implementable scientific results that will have measurable impact on society and the Earth's wellbeing. Aspects of openness, reproducibility and, thus, trust are fundamental in all examples.

- ***Overcome existing fragmentation of data and knowledge products in environmental sciences.*** Meeting the environmental challenges of our time requires interdisciplinary and increasingly transdisciplinary research practices involving many domains of independent research that lead to a proliferation of data sources, standards and tools. With this the need of integrated and seamless access not only to climate model data, but also to other environmental sciences (e.g. hydrology, agriculture), social, economic and biological data at a whole range of time and space scales is becoming critical. This is particularly acute for international programmes such as those of the Belmont Forum. Barriers to addressing these challenges must be overcome. Increasing linkages are needed, using `translational tools' among a diverse suite of data and knowledge products, to overcome this fragmentation and enable discovery processes in an interdisciplinary/transdisciplinary context involving, for example, environmental sciences, health, social and economic sciences and public authorities. Innovative solutions, using the best practice of both public and private sectors, are needed to overcome these issues so multi-source data can inform decision support systems, assess the benefits of new technologies or management, changes in climate, and trade-offs between productivity gains and environmental risks. There are quality control challenges, to be balanced with the rates of ingest and annotation, and this need exposes a major gap in transforming subsystem knowledge into holistic knowledge, e.g., from the potential value of data collected in single disciplinary experiments to something that can be used across multiple disciplines. At the same

[1] These examples are designed to help illustrate, for GPC members, the intent of this proposed CRA Call; this is *not* a prescriptive list of suggested projects.

time, data synthesized from different sources often cannot easily be transferred back to knowledge on subsystems. In addition, many existing global data centres do not handle all the types of regional data required, and typically selected variables are only provided at annual, monthly and daily frequencies. This area will address these challenges.

- *Create and sustain national and transnational foundation for multidisciplinary and multi-source data integration and analysis systems.* Rapid evolution of data and computing infrastructure capacities and capabilities, together with high-speed networking and data movement protocols, enable innovative data integration and analysis–and their cost-effective engineering–of very large volumes and diversity of multi-source data from research experiments, observational and monitoring systems and extreme scale numerical simulations. Multi-disciplinary and multi-source data integration and analysis platforms, supporting *federated data and software systems*, are becoming critical today for extracting and sharing new research-based knowledge and services, as well as distilling information in support to decision making and adaptation policies, bridging the gap between government, general public on environmental issues. Such platforms and federated systems must bring data users/producers/stewards, together with data scientists. Moreover, they will also need to accommodate different space and time scales, diversity and complexity of data while facilitating data discovery in a resource-efficient way in terms of data storage, retrieval, manipulation, integration, and diversity of analysis at each stage of the complex workflows. These federated systems must also be capable of dealing with multi-source data and facilitate model inter-comparison as well as cross-model comparison, minimizing the time spent finding, using, and storing the data. As such they can provide an transdisciplinary framework where scientific knowledge and discovery can transcend disciplines. This is particularly important for problems addressed by the Belmont Forum. Data and software developed in pursuit of a given experiments can also be repurposed for other experiments, opening the path to cross-fertilization between disciplines. While such technological advances are increasingly accessible to data scientists, technological and procedural barriers remain which, if addressed, can ensure more domain scientists find and use these data sets in their research, and applications are developed that can ensure these data inform management and policy decisions. Building on existing national investments whilst expanding and federating these internationally can help partners leverage different expertise in order to shape common services through international coordination across communities.

- *Improve access and rapid analysis for disaster monitoring and early warning systems*. Anticipating and responding to major disasters (earthquakes, tsunamis, landslides, climate, water, flood, agriculture and land-use, health, etc.) requires seamless access to data and the capacity for rapid analysis that can inform timely decisions. Disaster monitoring and early warning systems require minimizing the time spent finding and accessing data, data movement, data interoperability, rapid, continuous and near real-time streaming analysis of multi-disciplinary, and multi-source data that are generated from high-performance simulations and monitoring systems, all through distilling and collating information into forms that can be routinely used in decision making. We also need to address imperfect and incomplete data, together with the associated need to quantify uncertainty. At the same time, the trans-disciplinarity of disaster monitoring and early warning systems present considerable challenges for collaboration, methodological and technological including, for socio-economic security and resilience, key sectors responding to or affected by major disasters. This is particularly acute in the international context represented by the Belmont Forum, but also means the Belmont Forum is uniquely placed to play a leading role. New innovative disaster-information platforms, taking advantage full advantage of most advanced developments in observation analysis, data-aware e-infrastructures, high-performance computing and networking technologies while rebalancing attention to the full path of multi-source data use through adaptation knowledge services can have sustained impact in this area.

- *Enable high-end data assimilation through improving the convergence of high-performance computing (HPC) and high-end data analytics (HDA).* Data assimilation combines real world data from both transnational, multi-source observations and model simulation outputs in an optimal way to provide a more complete and coherent description of the environmental system and thus improve our predictive capabilities for distilling information useful to stakeholders ranging from government administrations to the general public about impending high-magnitude environmental events (e.g. climate and weather, flood, tsunami, and landslides forecasting applications). Today, data assimilation is a prime example of the immediate utility of the convergence between HDA and HPC through complex and data-intensive workflows that challenge the capabilities of data and computing e-infrastructures. For each regional weather forecast, data must be found, gathered and processed from multiple sources, including satellites, dropsondes, weather stations and buoys, current and historical observations, and simulations of the Earth's physical patterns and processes. All of this information is then input into complex, nonlinear models that simulate weather patterns that are likely to occur and produce outputs that can be compared against observations using data assimilation. The differences between characteristic features of the predicted and observed data are inputs to the assimilation engine. With the availability of increasing computing power and networking, together with the advances of stochastic optimization and machine learning techniques, uncertainty quantification can be estimated using an ensemble consisting of a large number of independent simulations. As computing and sensing technologies advance rapidly, projects devoting efforts to innovative and rapid statistical, multi-source data assimilation methods and tools orchestrating HDA and big HPC simulations across networked data and computing infrastructures are today critically needed to take full advantage of the combination of most advanced developments in observation analysis, high-performance computing and networking technologies and revolutionize Big Data rapid data assimilation to be used for warnings about environmental and climate hazards. Many agencies are involved in data assimilation, but the international coordination of methodologies is only beginning. The Belmont Forum can thus play a leading role.

- ***Broaden and accelerate international Data Model Intercomparison (DMI) experiences.*** Data model intercomparison (DMI)–and cross-model comparisons–has become standard research practice in environmental sciences (e.g. coupled climate and agriculture crop models). This sort of data sharing critically depends on global data and software infrastructures supporting federated systems, e.g. the "vast machine" of Edwards (2010)[2]. The increasing resolution and complexity of comprehensive coupled multi-system environmental models is rapidly leading international modeling centers and research groups around the world to run very large simulations–comprising more and more versions of these models–and thus to a growing volume and list of necessary and desired model outputs for related decision-making and operational services (e.g. climate services). In such a context, running a multi-model data analysis experiment is very challenging. This is particularly acute for international programs such as the Belmont Forum. Barriers to addressing the multiple challenges that DMIs are facing need to be overcome to enable the availability of a large amount of data related to multiple environmental models simulations together with persistent organized sets of scientific data management and movement protocols and tools for high-performance large-scale data analytics. This includes among others transnational multi-source data access and sharing, end-to-end workflow management across federated and flexible data and computing infrastructures, interoperability with the communities eco-system, data replication and versioning, curation, reusability and archiving. DMI experiments push also for improved formats, standards and model documentation due to the growing complexity of these models and in recognition that an increasing number of users want access to DMI data. This must also enable a substantial data assembly effort, focused on gathering and converting observations and reanalysis products into accessible formats for use in model evaluation. Barriers are also procedural as DMI campaigns require sharing of hardware and software expertise and exchanges between model users and the research community, minimizing duplication of efforts and reducing operational and computational resource demand. The new Belmont Forum call will broaden and accelerate DMI experiences, and as such facilitate systematic model evaluations as part of subsequent assessments. Developing faith in models and building confidence is crucial when the information extraction is being used immediately for important decisions under stressful pressure of emergency response or for revising planning of habitation and businesses or adapting agriculture, land use or water management.

[2] Edwards, P.N., A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming. Cambridge MA; MIT Press, 518 pp., 2010, ISBN: 978-0-262-01392-5.

- ***Optimize the creating and sharing of methods for interdisciplinary exploitation of data.*** There are a myriad ways of exploiting data with more being developed by scientists every day; for each of these there may be dozens of algorithms, each of which may have many implementations to suit different data and different computational contexts. Each of these approaches needs to be calibrated, validated, measured and described. They often need modification – for example, in urgent-computing contexts they need to trade accuracy for speed, whereas, in monitoring contexts, they need incremental algorithms to process continuous streams. This "methods and tools" research is a vital intellectual endeavour; made more challenging if every step, from data to well-presented compelling evidence, is to integrate well together and be easy to use. Innovative projects that catalyse transnational collaborative development of these new methods, software and infrastructure capabilities and identify what to share–-through strong and mutually-dependent collaboration among domain scientists, data and computing scientists and engineers, and infrastructure providers– are therefore critically needed both because such development is costly in skilled human labour, and because trust in results will depend on shared experience of their efficacy.

- ***Increase science reproducibility and trust through improved e-infrastructure to track and share research outputs from specific projects.*** The typical research lifecycle starts with the discovery of resources –data and computing– followed by their acquisition. The actual conduct of the work comes next, followed by the publication of resulting papers, data and methods – all of which benefit from curation. Reproducibility of scientific results and trust in these results is important for *researchers*, *publishers*, *funding agencies and stakeholders*, and need to be incorporated within research practices all through the knowledge cycle. This is both a cultural as well as a technological problem, and so solutions should include common standards, policies and tools designed to facilitate the tracking (e.g., appropriate common identifiers), archiving, reuse and visualisation of scientific results which are published and archived. With this comes the need to link agreed identifiers for funded research projects and web accessible versions of publications accurately to new digital objects that bundle together data, experimental descriptions, algorithms, models, software, workflows, provenance information and results in a standardized way. This is particularly acute in the international and transdisciplinary context of the Belmont Forum. This process will also enable other researchers, building on the work that has been done previously, to progress faster than they would have done otherwise, improving communication, sharing and reuse within and across disciplines. The three stakeholder groups would benefit as all are motivated by appropriate metrics and credit. However, this raises new issues of privacy and intellectual property. When data is integrated from multiple sources, be it by distributed query, workflow or mashup, the question of the licence on the result is an issue requiring progress in legal understanding and practice.This also requires strong commitments and stewardship of the research communities, together with mutually dependent collaborations across research communities, data scientists, ICT researchers, data providers, funding agencies and publishers.

- ***Improved trust in the security of international and transdisciplinary data and results.*** Transdisciplinary and international researchers in particular must be able to trust data and scientific results if they are to use them as they will often not know the full scientific context in which the data and results were collected, particularly if they normally work in a different country or a different scientific domain. Tools for demonstrating integrity, and tracing what changes have been made to data and results, need to be developed for this area of transdisciplinary and international research, particularly given the social concerns such as privacy which limit purely technical solutions. The solutions proposed must be acceptable to domain scientists, across the areas of interest to the Belmont Forum, which presents ideal test cases for this work.

- ***Implement shared and trusted multidisciplinary and multi-source open data management policies in environmental sciences.*** Implementing common data management policies and principles requires data provenance system and data citation system that rewards the provision of excellent open data. One of the least developed aspects necessary to ensure excellent open data is the concept of the scientific quality of the data, which can be defined in terms of accuracy, precision, uncertainty, validity and suitability for use (fitness for purpose). Considerable challenges remain in harmonizing (across disciplines) and standardizing (within disciplines) how scientific quality is assessed and documented, how the corresponding metadata is formatted, and how the associated information might be published and made traceable throughout the data lifecycle. Finally, a robust monitoring system needs to be established to track all research data generated by public funding, how it is curated, maintained and preserved for future generations as well as how it is re-used. A first step could be the development of a pilot discovery catalogue that harvests metadata from datasets produced by Belmont Forum projects and that is flexible enough to account for the differences across Belmont Forum funding agencies.

- ***Develop new data or metadata interoperability enabling data and model intercomparisons in transnational and interdisciplinary environmental research***. New types of data–from observations to results of model simulations and data analysis–and data collection approaches may not be accessible and well-characterized by existing data protocols, and data and metadata structures. The development of agreed data protocols and standards for data types related to environmental change research can help pulling data from various transnational sources and standardize data and ontology sharing, together with provenance information, to facilitate transnational data use within and across disciplines and eventually the development of appropriate federated transnational archives.

- ***Lower barriers to open science practices, especially in the context of citizen science***. Citizen science approaches to tackling research questions range across a broad spectrum from co-designing research questions and methods to low-investment contributions to data collection. Open science practices (encompassing everything from open notebooks and preregistration of study designs to open access publications and open data) also exist on a spectrum. There are few projects currently incorporating strong elements of both open science (or even just open data) and citizen science. Infrastructures are needed to facilitate the implementation of better open science practices along with engagement of citizen scientists (at whatever levels are appropriate to the environmental change problem at hand). Many Belmont Forum agencies are implementing citizen science projects, many of which are impressive; however, developing standards, estimates of accuracy of results, ways of feeding results back to participants, and sharing methods internationally, will benefit greatly from the international sharing of methods and best practices.