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

(a Concept Note for the upcoming Belmont Forum CRA)

Executive summary

The Belmont Forum has demonstrated it is possible to dramatically increase the impact of environmental change research through transnational cooperation among science agencies. The Belmont Forum is also ideally positioned to help the scientific community overcome critical technological and procedural barriers to optimizing the impact of research data collected across borders and disciplines. Many Belmont Forum members already have the capability to bring computer science and technology as well as large and complex data sets to bear on interdisciplinary science. However, no member has international best practice in all research areas, and the international nature of the Belmont Forum brings the needs of interdisciplinary science in this area into sharp relief. By working together on the highest priority problems, the Belmont Forum can establish international foundations for federated data integration and analysis systems with shared services, convene the best practices in public and private sectors, foster open data and open science stewardship among the science communities, including in related areas such as publishing, as well as encouraging data and cloud providers and others to adopt common standards and practices for the benefit of all.

This Concept Note provides the framework for the Belmont Forum Group of Program Coordinators (GPC) to develop a three-to-four year competitive Collaborative Research Action (CRA) funding call designed to provide targeted support to initiatives that are well-positioned to solve one or more of the methodological, technological, and/or procedural challenges currently facing environmental science researchers working with large, complex and multi-source data sets. Projects financed by this Call will bring domain scientists and information and communications technology (ICT) experts together to produce demonstrators of concrete, scalable solutions that are relevant to the larger scientific community and which accelerate the full path of data-use across existing data and computing infrastructures, as well as the knowledge life cycle, from data acquisition all the way to decision-making. The originality of the Call is its implementation as a "task force" requiring all funded projects to share results and participate in regular steering workshops (e.g., one per year), and to contribute to a knowledge hub. The resulting analyses will be used to inform the evolution of e-infrastructure investments, and will be reflected in subsequent iterative funding calls where further strategies may be identified or existing strategies may be enhanced.

The overall goal

The goal of this Call is to bring modern computational and analytical sciences and associated infrastructure providers to bear on answering some of the most pressing analytical, computational and data-related needs of transnational and interdisciplinary environmental change research, and to catalyse such research efficiently with international best practice. These environmental sciences needs concern both individual research questions as well as the broader research community. Past and recent Belmont Forum research has clearly illustrated these needs.

New scientific discoveries and socioeconomic innovation will emerge when tackling the exponential increases in diversity, volume and throughput of multidisciplinary data in environmental sciences. It is critically important to establish and enable interdisciplinary and transnational frameworks so that scientific knowledge can transcend disciplines and geographical borders. For the benefits of scientific knowledge to be realized by society, we may also need to combine knowledge originating from natural sciences, health, socioeconomics and humanities.

The Call emphasizes 'going the last mile' with data: not only uncovering the evidence that will support a scholarly observation, but also distilling and collating the evidence into forms that can be used routinely in research and decision-making so that the data ultimately increases the wisdom behind policy and action, i.e., using research to help people take appropriate action. For example[1] the freshwater supply of California, as in many other parts of the world, is threatened by continuation of river management policies that have not adapted to changes induced by global environmental change. Scientists and decision makers need now to use satellite and field data to measure the water stored as snow so that the changes in timing and volume of flow can be understood. Decision science is even more demanding of data-intensive methods and technologies, as urgency forces the use of imperfect and incomplete data.

Many of the individual Belmont Forum agencies already sponsor research that brings computational and environmental sciences together. However, this has revealed gaps and barriers as well as the increasing need to accelerate the full path of interdisciplinary data use and enhance reproducibility and trust in scientific results. The aim of this Call is to bring these efforts together, driven by the international and trans-disciplinary nature of Belmont Forum research that brings the urgency of the need into sharp relief.

Projects responding to this Call should provide concrete and innovative demonstrators toward solutions to pressing science-driven issues and to co-evolve research practices with emerging

[1] From: Jeff Dozier and William B Gail. The emerging science of environmental applications. In: Tony Hey, Stewart Tansley, and Kristin Tolle (editors), The Fourth Paradigm: Data-Intensive Scientific Discovery, pages 13–19. Microsoft, 2009.

new methods, Information and Communication technologies, and their supporting software, for the benefit of all while sharing existing best practices.

Hence, the SEI CRA seeks to provide targeted support to initiatives that are well-positioned to address one or more of the methodological, technological, and/or procedural[2] challenges currently facing environmental science researchers working with large, complex and multi-source data sets. Proposals should thus be designed to provide focused, science-driven solutions that may serve as demonstrators to support the broader scientific and informatics communities that may face similar challenges.

The Concept Note context

A team led by ANR-France and co-led by JST-Japan and MoST-Chinese Taipei elaborated this Concept Note in coordination with the secretariat of the Belmont Forum e-Infrastructures and Data Management (e-I&DM) CRA[3]. The SEI Call will ultimately be defined and proposed by a Group of Program Coordinators (GPC) comprising representatives of the funders participating in the Call.

What are the specific objectives of the SEI Call?

The SEI Call will be designed to:

- Address real world, *researcher-experienced problems* arising from well-identified methodological and technological barriers, or procedural issues, to enable scientific discovery and accelerate the full-path of transnational and interdisciplinary data use[4] through the publication and distillation of information that impacts decision-making.
- Provide funding to new, innovative projects involving mutually-dependent, science-driven and solution-driven *collaboration among domain and data scientists, ICT researchers and engineers, and infrastructure providers.*
- Co-evolve research practices with new analytic methods, infrastructure capabilities and technologies, and their supporting software with the aim of contributing to better use and reuse of data and information resulting from peer-reviewed research on global environmental change and sustainable development issues.
- Engender collaborative development, implementation and sharing of *new methods, tools,*

[2] Procedural issues may be organizational, legal, security, and/or policy related, with respect to the potential benefits and limitations of open science, from data collection and management through the publication of results. Procedural issues may also influence the capacity of domain scientists to effectively use computing and data infrastructures, as well as to organize the ongoing integrity, accessibility, interoperability, and reuse of data (data stewardship).

[3] Belmont Forum e-Infrastructures and Data Management Project (e-I&DM) Collaborative Research Action (CRA)

[4] The full-path of data use extends from data capture, data access and management, data movement, data analysis and modeling, data provenance through data and model inter-comparison.

software, protocols and mechanisms for interdisciplinary and transnational use of complex and diverse multi-source data in order to ensure that project results can be adopted in the larger data and domain science communities in a cost-effective manner.

- Deliver and share extended and inclusive data-intensive demonstrators and pilots toward solutions that are scalable via application to other contexts, and that could have sustained impact on research practices.
- Support open data and open science through innovation that enhances data quality, data citation, data sharing and reuse within and across disciplines, and that enables reproducible science.
- Enable and accelerate multi-source, transnational *data and model intercomparison* (DMIP) addressing environmental and socio-economic challenges related to global change and sustainable development issues.
- Develop *international collaboration* for innovative and potentially disruptive data and computing infrastructure capabilities building on federations of existing infrastructures,

  shared services, data analytic methods and software that *accelerate the rates at which information is gleaned* from multi-source and multi-disciplinary data.

A set of examples[5] is provided in the Annex solely to illustrate some of the barriers and issues that might possibly be addressed by this new CRA funding Call.

What are the key features of the proposed SEI Call?

The SEI Call will follow the standard Belmont Forum procedures for international, collaborative development of the Call across all interested agencies, and the format and timing of the Call announcement and review. In addition, this Call will include several innovative features:

I. The Call will be implemented as a "task force" (for example, see the Internet Engineering Task Force [IETF®]) including steering workshops, and a scientific hub[6] for sharing across the funded projects methods and their software implementation, good practices intended to deliver priority recommendations to the Belmont Forum, in particular for specific training and education needs and subsequent iterative funding calls where further strategies may be identified or existing strategies may be enhanced.
II. Projects will be selected competitively, but each must be open to collaboration and sharing with the other selected projects.

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

[6] The definition, implementation and the operation of the scientific hub shall be defined in the Call implementation plan by the founding members of the Call, involving possible in-kind contribution, and provided to the Theme Program Office that will manage the steering activities of the Call.

III. The prerequisite for any proposal is an interdisciplinary, science-driven approach with strong, mutually-dependent collaboration among domain, data and computing scientists, together with infrastructure providers, aimed at providing methods and ICT solutions that will enable and impact environmental change and sustainable development research and the reuse of data across disciplines.
IV. Projects must be transnational with collaborators from at least 3 different countries represented in the Call funding partners (3 of whom must be Belmont Forum members). Researchers not covered by participating funding agencies are also eligible to join projects using their own or other source(s) of funding.
V. Projects should focus on a science-driven problem covering one or more of the above objectives, and demonstrate how solving this problem actually accelerates the full path of data use and enables further scientific discovery.
VI. Projects must provide clear and demonstrable deliverables that could be adopted beyond the project by broader science communities in a cost-effective manner and with sustained impact.
VII. Projects must demonstrate the adoption and implementation of open data and how this will support and contribute to the open science principles (see below) and to comply with the Belmont Forum data policies.
VIII. The Call provides an agile implementation[7] of a 3 to 4-year funding track. Coordination between and among projects will be expected through their participation in regular workshops coordinated during the Call.

Who should participate?

This Call targets a combination of environmental change domain scientists and computer and data scientists aiming to accelerate the rates at which information is gleaned from multi-source and multi-type data, optimizing the impact of transnational and interdisciplinary environmental change research and ensuring that a broader scientific community benefits from the new and potentially disruptive solutions identified.

Various types of actors are encouraged to participate: public research organizations (government, universities, etc.), private data-related companies, foundations, non-government organizations, infrastructure providers, publishers, public-private partnerships, etc.

Public-private sector partnerships for co-design and co-development are encouraged and may include, for example:

- in-kind and service providers - e.g., cloud, data certification, repositories;
- developers of data access, security, analysis and management tools and libraries;

[7] Collaborative and iterative feedback and follow-up leading to continuous learning (from end users/stakeholders) and improvement of the projects' objectives.

- end-users of research results - e.g., insurance sector researchers that publish their work in the refereed literature, publishers, etc.

In the case of partnerships with the private sector, all results and products directly resulting from the funded projects will remain in the public/research domain.

The specific details of these partnerships are subject to national policies and constraints. Details will be provided in the official Call text and its annexes.

Project typology

The Call will provide 3 to 4-year funding for new transnational and interdisciplinary projects addressing well-identified and potentially ground-breaking and science-driven solutions. Projects should be:

Science-driven. Responding proposals should be driven by the objective of developing innovative solutions that enable and enhance interdisciplinary environmental change research while addressing tangible research-driven objectives. In other words, proposers will need to clearly explain how the project will specifically benefit, serve and/or accelerate data-driven, scientific research and discovery and lead to concrete results, and that the project outcomes could be applied in other scientific areas.

Collaborative. Projects should involve collaboration between and among domain scientists and computer and data scientists, and possibly infrastructure providers, to solve well-identified, experience-based issues and barriers in analytic methods and data-aware technologies.

Scalable. Solutions provided by responding proposals are expected to be scalable so that the larger domain and data science communities can benefit, and therefore have a sustained impact on research practices.

End-to-end. The projects should accelerate the full path of data use across existing data and computing infrastructures, and the knowledge cycle in relation to Belmont Forum research challenges. This extends from transnational, multi-source data management and access, including data interoperability, to multi-source data analytic and modelling methods for the extraction of new information that impacts decision-making in the Belmont Forum-related challenges.

Interdisciplinary. Great premium will be placed on research-driven, problem-solving strategies with an emphasis on the sharing within and across disciplines of: (1) data, data science methods, computing and data technologies; (2) multi-source and transnational data and model inter-comparison; and (3) new sharable object representations—integrating data, provenance

information, workflows and results—together with accurate identifiers enabling reproducible research through sharing and reuse.

Annex 1 provides examples of some of the immediate challenges that might be addressed with this new CRA Call and of interest to the Call funders. Further details will be provided in the official Call text.

Open Science requirements

The Belmont Forum Data Policy and Principles state that data should be:

- Discoverable through catalogues and search engines.
- Accessible as open data by default, and made available with minimum time delay.
- Understandable in a way that allows researchers—including those outside the discipline of origin—to use them.
- Manageable and protected from loss for future use in sustainable, trustworthy repositories.

These principles have been designed to encompass the objectives of interoperability and reusability, thus aligning with the FAIR (Findable, Accessible, Interoperable, Reusable) principles.[8]

Open Science, in addition, implies that all scientific results resulting from projects that are funded by public resources should be published in (as far as possible) open access journals or otherwise made publicly available.

All of the above are elements of a Data Management Plan that is mandatory in the Call and that will be monitored by the Data Policy and Planning Action Theme of the Belmont Forum e-I&DM CRA.

For further information…

A scoping workshop, in preparation of this Call, was held in Paris, France, in November 2016. The workshop projects and presentations can be consulted here: http://bfe-inf.org/info/eidm-scoping-workshop.

For more information and any questions directly related to this Concept Note and the upcoming Call, please contact Jean-Pierre Vilotte (jean-pierre.vilotte@agencerecherche.fr)

[8] See https://www.nature.com/articles/sdata201618 and https://www.force11.org/group/fairgroup/fairprinciples