Community Edition

10th August 2015

A Place to Stand: e-Infrastructures and Data Management for Global Change Research

Belmont Forum e-Infrastructures & Data Management Community Strategy and Implementation Plan

"Give me a place to stand, and I will move the world"
- Archimedes

Steering Committee, Belmont Forum e-Infrastructures and Data Management Collaborative Research Action

Lee Allison, Arizona Geological Survey (Co-Chair)
Robert Gurney OBE, University of Reading (Co-Chair)

TABLE OF CONTENTS

PREFACE

Global environmental change is considered to be one of the most pervasive concerns of the 21st century. Scientists throughout the world are independently undertaking research to determine the nature and extent of these changes, and their impacts on humans and the environment. Global change research enables scientists to understand and predict how our planet functions and evolves and to investigate responses to those changes.

> ***"Why should we care? Because, just as the World Wide Web has transformed our lives and economies, so this new data wave will matter eventually to every one of us, scientist or not."***
> *- The Data Harvest, RDA Europe, 2014*

This research increasingly requires integrating large amounts of diverse data across scientific disciplines to deliver the policy-relevant and decision-focused knowledge that societies require to respond and adapt to global environmental change and extreme hazards, to manage natural resources responsibly, to grow our economies, and to limit or even escape the effects of poverty. To carry out this research, data need to be discoverable, accessible, usable, curated and preserved for the long-term. This needs to be done within a supporting data intensive *e-infrastructure* framework that enables data exploitation, and that evolves in response to research needs and technological innovation. Without such data and the supporting e-infrastructure, policy makers and scientists will be forced to feel our way into the future without the benefit of new scientific understanding, unfocused and ill-prepared.

The Belmont Forum seeks to establish a cooperative approach to developing **sustainable practices within the global change research community** for data discovery, management and curation. The goal is to streamline the dissemination of global environmental change information and maximize the opportunities for effective action.

This edition is a public version of the report that was submitted to the Belmont Forum Principals but with some added context for the general reader about the Belmont Forum and the motivation for the report.

About the Belmont Forum

Established in 2009, the Belmont Forum is comprised of the world's major funding agencies of global environmental change research and international science councils. Guided by the Belmont Challenge, which aims:

"to deliver knowledge needed for action to avoid and adapt to detrimental environmental change including extreme hazardous events",

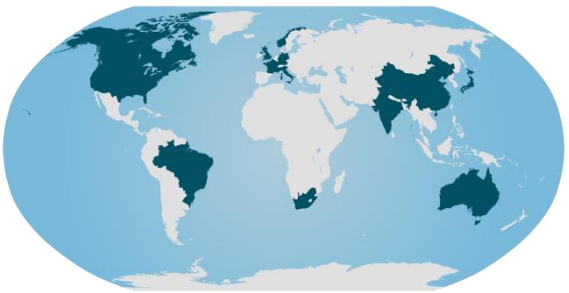

Figure 1: Belmont Forum Member countries

the Belmont Forum serves as a round table for these agencies to collectively address issues related to global environmental change. Initially, priority focus has included coastal vulnerability, freshwater security, ecosystem services, carbon budgets and most vulnerable societies.

To meet the goals of the Belmont Challenge, the Belmont Forum coordinates funding for Collaborative Research Actions (CRAs), which are high-priority research activities designed to improve the way funding agencies collaborate with each other and develop opportunities for multi-national research. To date, the Belmont Forum has funded five CRAs, with several more proposed or in progress. This CRA is the first to deliver its Community driven report and recommendations. For detailed information about the Belmont Forum, the Belmont Challenge and CRAs, visit www.igfagcr.org/.

Figure 2: Belmont Forum and International Group of Funding Agencies (IGFA) members (now in the process of merging)

E-Infrastructures and Data Management Collaborative Research Action

Accurate and reproducible science requires comprehensive and verifiable data that are appropriately documented and accessible. As researchers strive to understand the vast and varied systems comprising the global environment, a large factor determining their success is access to robust and reliable data as a foundation and reference point for their own conclusions. The Belmont Forum initiated the E-Infrastructures and Data Management CRA to collectively develop achievable and sustainable e-infrastructures and data management practices in recognition that:

> ***…the need to address global environmental challenges requires a more coordinated approach to the planning, implementation, and management of data, analytics and e-infrastructures through international collaboration.***
>
> ***— Belmont Forum, New Delhi, February 2013***

This Report is the result of activities conducted over an 18-month period by an international Assembly of more than 120 domain scientists, computer scientists, information scientists, social scientists and legal scholars. Their task was to survey the state of current practices and establish recommendations on how the Belmont Forum can leverage existing resources and investments to foster a more coordinated, holistic and sustainable approach to the funding and support of open and effective data management practices. The Assembly was guided by an international Steering Committee which consisted of experts from research and user communities from participating Belmont Forum member countries. Members of the Steering Committee were responsible for leading one or more Assembly working groups (Work Packages) in order to collectively assess existing international e-infrastructure capabilities, identify gaps and overlaps, prioritize challenges, and provide recommendations on how to best address the Belmont Challenge. Logistical and administrative support was provided by a joint US-UK supported Secretariat.

E-Infrastructures and Data Management CRA Members

The main sections of this Report were written by the project Secretariat with guidance and significant effort from the Steering Committee, with review and edits from the Assembly. This final report is a synthesis of 1) comprehensive reports by each Work Package on the state of the art, barriers, gaps and best practices, 2) Steering Committee contributions from a series of in-person and virtual meetings, and 3) feedback from meetings of national delegations of the participating Belmont Forum countries.

Figure 3: Some of the Belmont Forum e-Infrastructures and Data Management CRA members

This report is intended to prioritize actions best suited for the Belmont Forum collaboratively to address *interoperability* and organizational challenges in data management and e-infrastructure, and to identify existing national and international initiatives which demonstrate good practice to create a global momentum toward thoughtful data management. Therefore, this report aims to:

- Identify strategic science policies, outlining what can be done better, in a multilateral way, to support global change research;
- Clearly express global e-infrastructure needs, barriers and gaps;
- Inform stakeholders;
- Prioritize actions to address interoperability challenges

EXECUTIVE SUMMARY

An e-infrastructure that supports data-intensive, multi-disciplinary research is needed to facilitate new discoveries and accelerate the pace of science to address 21st century global change challenges. Data discovery, access, sharing and interoperability collectively form core elements of an emerging shared vision of e-infrastructure for scientific discovery. These elements further depend on building relationships among data sets, people, systems, organizations and networks. However, the pace and breadth of change in data and information management across the *data lifecycle* means that no one country or institution can unilaterally provide the leadership and resources required to use data and information effectively, or to establish and maintain the relationships needed to support a coordinated, global e-infrastructure.

The Belmont Forum represents many of the world's largest and most influential funders of environmental and social science research. It is uniquely capable of catalyzing international collaboration and leveraging existing national programs to effectively initiate and guide best practice in data stewardship, data sharing and e-infrastructure development to meet global change research needs. Furthermore, alignment of international and cross-domain efforts in interoperability will promote new interdisciplinary and international scientific understanding relevant to the Belmont Forum research agenda. As such, ***the Belmont Forum is ideally poised to play a vital and transformative leadership role in establishing a sustained human and technical international data e-infrastructure to support global change research.*** This *Community Strategy and Implementation Plan* (CSIP) proposes an initial path forward.

Recommendations

The Belmont Forum is urged to adopt the overarching and synergistic recommendations listed below, through its unique role in global research collaboration, to: fill critical global e-infrastructure gaps; improve data management and exploitation; coordinate and integrate disparate organizational and technical elements; share best practices; and foster new *data literacy* to enable actionable and societally beneficial science. These recommendations have the potential to transform the way data are used and research is conducted by accelerating discovery, increasing the value of research in decision-making, and catalyzing changes throughout the economy and society that are of value to all citizens.

The five recommendations are:

1. **Adopt Data Principles** that establish a global, interoperable e-infrastructure with cost-effective solutions to widen access to data and ensure its proper management and long-term preservation. Researchers should be aware of, and plan for, the costs of data intensive research.
2. **Foster communication, collaboration and coordination** between the wider research community and the Belmont Forum, and across Belmont Forum projects through a Data and e-Infrastructure Coordination Office established within a Belmont Forum Secretariat.

3. **Promote effective data planning and stewardship** in all research funded by Belmont Forum agencies to enable harmonization of the *e-infrastructure data layer* through enhanced project data planning, monitoring, review and sharing.
4. **Determine international and community best practice** to inform e-infrastructure policy for all Belmont Forum research, in harmony with evolving research practices and technologies and their interactions, through identification and analysis of cross-disciplinary research case studies.
5. **Support the development of a cross-disciplinary training curriculum** to expand human capacity in technology and data-intensive analysis methods for global change research, and increase the number of scientists with cross-cutting skills and experience in best practice.

MOTIVATION

A New Data Literacy for the 21st Century

The United Nations noted that the world needs a new data literacy that enables actionable and socially-beneficial science to address environmental change affecting disaster mitigation, resilience, water and other natural resources.[1] Broader and more effective development of best practice in data stewardship, sharing and cross-disciplinary use are pillars of the new data literacy and the basis of *Open Science* and, more generally, of the direction of science itself. Global access to data will change the ways we address environmental change problems and also change our behavior; mastery in the management and exploitation of data is key to successful collaboration and future research.

> ***"There is no turning back the clock on our interconnected world, but we could jeopardize its benefits if we fail to invest in a trusted data environment."***
> *- Ellen Richey, Chief Enterprise Risk Officer, Visa, USA, in WEF Blog on Big Data*

Unique Challenges in Global Change Research

Global change research is a crucible for shaping e-infrastructure technologies and research practices. Free and open exchange of data, methods and results, as well as effective data stewardship, are central to advancing scientific enquiry in all fields but there are particular challenges and needs in cross-disciplinary research areas. Challenging multi-disciplinary research questions relating to the Earth system span physical (e.g. atmosphere, land, and oceans), political, social and geographical boundaries, requiring data and information to be interoperable and exchangeable worldwide. Global change research also integrates diverse observations, data-intensive analytical methods and numerical models across numerous scientific domains. It requires extensive data storage and movements, including emerging capacities in *cloud computing* and *High Performance Computing*. In addition, there is a need to preserve historical, often "small" and disparate data, as much of global change research relies on observations that by definition cannot be repeated. Both the public and

[1] A World That Counts: Mobilizing The Data Revolution for Sustainable Development. undatarevolution.org/report/.

commerce have a high level of interest in the results, leading to an increasing demand for veracity, dissemination and citizen involvement.


Subsistence Farming in Yunan Province, China

Importance of Overcoming Historical Barriers to Interoperability

Major regional, national and international e-infrastructure efforts[2] have noted that cultural, social and organizational barriers to global data sharing and interoperability generally exceed technical barriers. These non-technical aspects are easily overlooked or considered outside the scope of domain and of information and communication science and technology programs. Funding strategies by research agencies have also inadvertently bolstered these barriers by supporting investigator- or discipline-generated projects that are generally disconnected from each other and are typically independent of an overarching, integrated framework that would contribute to a coordinated e-infrastructure. Similarly, policy has often focused independently on particular segments of the data lifecycle (such as data acquisition, storage and distribution or data-intensive High Performance simulation) whereas a policy which bridges the whole data lifecycle is required for a healthy data-intensive e-infrastructure environment. Thus, the emphasis in this report is to integrate across the technical and non-technical aspects of interoperable data and e-infrastructure.

Reproducibility in Science

In October 2014, the Belmont Forum Principals requested that this CRA consider issues regarding reproducibility in science. Elements of reproducibility underpin all science, including global change research. They include: reuse of data and code; need for data repositories and sharing platforms; standards required for sharing code and data effectively and accurately; citation, *provenance*, *metadata*, tools and incentive mechanisms; capture and sharing of workflows; and ensuring domain-specific statistical reproducibility in the computational and data science software stack. Accurate capture and free exchange of data and information is inherent in this. Reproducibility is thus not drawn out separately in this report but is interwoven into its

[2] COOPEUS, RDA, ICSU-WDS, DataONE, DIAS, ESIP, EarthCube, GBIF, GEOSS, iCORDI, INSPIRE and OneGeology.

conclusions and recommendations. The term "reliability" of data is emerging as a possible alternative descriptor of the issues involved in reproducibility of science.

FINDINGS AND RECOMMENDATIONS

Our Vision

Our vision is of high quality, reliable and multidisciplinary global change research enabled by a sustained human and technical, internationally coordinated and data-intensive e-infrastructure able to process a continuous increase in the diversity and volume of data generated. In such a research-driven e-infrastructure, data should be discoverable, reusable, open and accessible by default as far as possible. In addition, the data's fitness-for-purpose should be assessed using transparent metadata relating to trustworthiness and quality. To realize this vision and maximize the return on public investments in research, all stakeholders need appropriate incentives to contribute to and support this vision. ***The Belmont Forum can blaze a path towards achieving this vision by implementing the recommendations outlined below.***

Adopt Data Principles

Adopting the ***five data principles*** listed below, through the authority and reputation of the Belmont Forum, will help establish a global and interoperable e-infrastructure to widen access to data and ensure its long-term preservation in global change research.

Research data must be:

1. **Discoverable** through catalogues and search engines, with data access and use conditions, including licenses, clearly indicated. Data should have appropriate persistent, unique and resolvable identifiers.
2. **Accessible** by default, and made available with minimum time delay, except where international and national policies or legislation preclude the sharing of data as *Open Data*. Data sources should always be cited.
3. **Understandable and interoperable** in a way that allows researchers, including those outside the discipline of origin, to use them. Preference should be given to non-proprietary international and community standards via data e-infrastructures that facilitate access, use and interpretation of data. Data must also be reusable and thus require proper contextual information and metadata, including provenance, quality and uncertainty indicators. Provision should be made for multiple languages.

4. **Manageable** and protected from loss for future use in sustainable, trustworthy repositories with data management policies and plans for all data at the project and institutional levels. Metrics should be exploited to facilitate the ability to measure return on investment, and can be used to implement incentive schemes for researchers, as well as provide measures of data quality.
5. **Supported** by a highly skilled workforce and a broad-based training and education curriculum as an integral part of research programs.

The development of these principles was informed by data principles generated and recommended by many international programs, such as the G8. These principles underpin the recommendations in this report as they inform the nature of the data plans and help identify best practice.

Foster Communication, Collaboration and Coordination

An appropriate organizational and community-building environment is necessary to: resolve barriers and gaps in global data sharing and interoperability; build relationships; distill information from data; and align incentives for effective and collaborative data management. Otherwise, the current trend of competing or conflicting technology development and agency policies will endure. While this work is, and will continue to be, undertaken largely in a national context, the Belmont Forum can place it into a global context by fostering the appropriate coordination and collaboration environment. ***The Belmont Forum can and must champion the organizational, community-building and technical framework needed to facilitate the international and interdisciplinary exchange of global change information through its member organizations, both individually and collectively.***

Promote Effective Data Planning and Stewardship

Communicating best practice in data and information stewardship and sharing will not only help to improve collaborative efforts but also reduce the associated risks and costs of data management. This involves: paying attention to the full lifecycle of data use and the rates at which information is gleaned from data; changing policies to promote better and more effective data planning; adopting data stewardship principles; and implementing incentives for their adoption, similar to the ways in which scientists are incentivized to publish research results. Establishing good practice is fundamental to improving data availability and interoperability. It will enable co-evolution of research needs with e-infrastructure, increase data usefulness, build trust among stakeholders, and reduce overall costs resulting from ineffective data management. ***The Belmont Forum is ideally positioned to achieve significant impact by collectively changing grant funding policies and reward systems to promote more effective data planning and stewardship.***

Determine International and Community Best Practice to Inform Belmont Forum Research e-Infrastructure Policy

Individual research domains successfully exchange best practice, either through scholarly publishing or increasingly through exchanging information via the Internet using a variety of mechanisms and applications. While there are beacons of good practice, there are inconsistencies in the exchange of information and the shaping and sharing of data-intensive e-infrastructure between nations and across domains and users. The rapid pace of change in technology and its adoption makes the normal development of good practice difficult and it is unclear whether the market will produce suitable solutions without intervention. Environmental and social sciences have a strong need to preserve and exchange information globally and all Belmont Forum members have examples of good practice to share. ***The Belmont Forum is uniquely placed to review worldwide and discipline-specific current practice and to foster best practice (in data sharing stewardship, analysis, modeling and workflows, and in the implementation of e-infrastructures) to promote efficiencies and trust in data and e-infrastructure solutions.***

Support the Development of a Cross-Disciplinary Training Curriculum to Build Capability

E-Infrastructures globally lack enough skilled people who understand data management and data intensive methods in environmental, social and health sciences, and in engineering to effectively drive this area forward. While training exists in a number of domains, it is frequently restrictive in scope. In addition, formal training is typically aimed at university students and early career researchers but there is a strong need for established scientists to become more data-enabled and data-proficient. Significant progress in building this capability can be achieved through cataloguing, accrediting and enhancing existing training efforts, filling critical gaps in a nascent global curriculum, and sharing methods for interdisciplinary and transdisciplinary exploitation of data. ***The Belmont Forum is well placed to stimulate new ways of thinking and working amongst distributed and diverse researchers, data and information scientists and data-enabled domain scientists, enabling them to better address global change research challenges.***

> ***"If you want to go fast, go alone.***
> ***If you want to go far, go together."***
> *- African proverb*

LEVERAGING THE POWER OF THE BELMONT FORUM

If the planet were a patient in a modern intensive care hospital unit, there would be a coordinated set of sophisticated monitors and instruments, rapid analysis and presentation of test results, a team of medical professionals coordinating diagnosis and treatment according to proven medical principles and best practices, and a set of available experts from different specialties drawing on the best available medical research and data. The Belmont Forum is in a unique position to develop key pieces of a comparable global e-infrastructure. It can act as a catalyst for promoting dialogue and collaboration, and leverage - but not replace - existing

national programs. It also provides a synergistic, top-down approach that complements bottom-up activities carried out by individual nations and organizations across the globe.

Implementation of these recommendations could include adopting internal actions and policies to align Belmont Forum efforts with external developments, influencing research investments judiciously, targeting limited resources where they are uniquely or best suited, or issuing funding calls (such as a networking or community-building action, a call to run a summer school or develop training materials, small-scale priming activities, large-scale research activities, or whatever is most appropriate to address the issue in question). For some actions, the Belmont Forum could identify that a CRA or invitation to tender would be the best funding mechanism to address an issue.

The challenges and opportunities in creating coordinated, global and interoperable e-infrastructure are complex but addressing them will result in tremendous benefits to stakeholders at all levels. These challenges are also clearly outside the ability of any single entity to attempt to control or implement, both in terms of resources and authority. Development of an e-infrastructure capable of supporting the existing and emerging global change research agenda has been, and will likely continue to be, organic with many aspects unpredictable and disruptive. It must therefore be agile and adaptable to meet changing research needs and technology development. Shared responsibilities are a key to success.

Shared Responsibilities

We have described the rationale for the Belmont Forum to undertake the recommended actions but have not discussed what the larger research and computing communities should do for Belmont Forum e-infrastructure and data management actions to be successful. Do individual Belmont Forum members take independent action? What should external entities and funding agencies do to support these activities? How does the Belmont Forum respond to external dynamics?

Globally, researchers and governments alike are recognizing the importance of data discovery, access, information sharing and interoperability. These collectively form core elements of an emerging shared vision of e-infrastructure for scientific discoveries, governance and resource management. There are numerous challenges to achieving these ambitious goals, many of which have been identified through existing Earth and related science informatics community initiatives. This broad, loosely-coupled community has identified many of the technical and social challenges to e-infrastructure but developing solutions that are adopted and collectively enhanced by the scientific community is still difficult. By building a cohesive international community committed to this e-infrastructure vision, the Belmont Forum can create opportunities for shared and more sustainable efforts toward removing barriers to interoperability on a global scale.

ACTIONS TO CATALYZE RECOMMENDATIONS


A climate for growth

Action Theme 1: Coordination Office

Foster communication, collaboration and coordination through the establishment of a Data and e-Infrastructure Coordination Office

The Belmont Forum should establish a Data and e-Infrastructure Coordination Office within the Belmont Forum Secretariat to foster communication, collaboration and coordination across Belmont Forum funded projects and to engage with the wider global e-infrastructure community. Its main tasks will be to support communication, collaboration and coordination among Belmont Forum-funded activities, e-infrastructure organizations, projects and experts; to promote greater awareness of the wider landscape of activity; to facilitate access to external resources; and to foster cooperation among and with other projects. The Belmont Forum should undertake this Action Theme because no other national or international bodies view this as part of their mandates or authorities, and there are no viable alternatives identified to-date for other bodies to assume responsibility.

The Belmont Forum should initiate the Coordination Office by appointing a Data e-Infrastructure Officer (DIO) and a Communication, Collaboration and Coordination Officer (CCCO), with staff support as needed, along with appointed "champions" from the other Action Themes (Data Planning, e-Infrastructure and Human Dimensions) to liaise with this Coordination Office.

Impact: This Action will enhance collaboration and general practice among Belmont Forum members and beyond to those institutions involved in global change research, data management and stewardship and e-infrastructure development. It could harmonize duplicative efforts and organizations, lessen volunteer fatigue, reduce redundancy and increase the impact of funding initiatives. Overall it should lead to a broader and more effective e-infrastructure integrated with network and computational elements.

Action Theme 1: Coordination Office
Near- and Medium-Term Actions (start within 0-5 years)

1. Establish (near-term) and provide support (medium-term) to a **Data Policy Advisory Board (DPAB) and Security Advisory Board (SAB)** to oversee relevant legal and security issues and contribute toward the development of the Enhanced Data Plan Template.
2. Oversee an initial (near-term) and ongoing (medium-term) **mapping of organizations and best practice activities**, to identify the people, projects, programs and organizations working toward data and e-infrastructure interoperability in global change research, the specific roles these individuals and groups are playing, and the best practices they employ.
3. Initiate the establishment (near-term) and management (medium-term) of a **Strategic Coordination Network (SCN)** of organizations and entities to identify collaborative strategies to address Belmont Forum objectives and actions to implement a more coordinated, holistic, and sustainable approach to the funding and support of global environmental change research. Develop and maintain mechanisms to foster communication and information sharing across these efforts in the medium to long-term.
4. **Oversee** the implementation, monitoring, and evaluation (near term) of **Belmont Forum e-infrastructure activities aligned around the other Action Themes in this report** (Data Planning, e-Infrastructure and Human Dimensions). Regularly report results, challenges and findings to the Belmont Forum and stakeholder communities (medium-term).
5. **Foster coordination** (near-term) **among Belmont Forum-funded projects to share and disseminate best practices** in e-infrastructure and data management, leverage work and results, and participate in case studies and workshops (medium- and long-term).
6. Provide an **auditing function for Data Plans submitted by Belmont Forum-funded projects** (medium-term) as described in the Data Planning Action Theme.
7. **Organize initial scoping workshops** to inform future case studies, human dimension capacity-building curriculum and data plans (see the e-Infrastructure Action Theme).
8. **Coordinate with external bodies** regarding any functions and tasks of the Coordination Office that are outsourced (near-term and medium-term).

Action Theme 2: Data Planning

Promote effective data planning and stewardship in all research funded by Belmont Forum agencies

This Action Theme proposes requiring Belmont-Forum funded researchers to provide and comply with an Enhanced Data Plan (EDP) overseen by a Data e-Infrastructure Officer. The primary goals are to enable harmonization of e-infrastructure through enhanced project data planning, monitoring, review and sharing. EDPs should cover the management of data, connection of the data to associated context (research project, resulting publications, researcher details, data provenance, etc.), facilitating the discoverability of the data, and ensuring that the data is capable of reuse and exploitation.

Impact: Establishing Belmont Forum good practice through requirements and monitoring under a common Data e-Infrastructure Officer will improve data availability and interoperability; enable co-evolution with e-infrastructure to ultimately help increase data usefulness; build trust among stakeholders; and reduce costs. Actions undertaken by the Belmont Forum to adopt and use community good practices for research data planning will ultimately benefit existing and future Belmont Forum-funded research:

- Improved data management and exploitation practices will enable international multi- and cross-disciplinary approaches needed to respond to the global change challenges as well as unlocking additional economic value from data holdings by facilitating data reuse.
- Recording and reviewing successful Data Plan implementation will, over time, lead to improved Belmont Forum data processes and advise the development of future e-infrastructures.
- Greater trust in the outputs of research and their applications will be facilitated because data will be more accessible and useful for validation of research results.
- Funder mandates can be a useful driver of compliance, by supporting coordinated engagement of various stakeholders across the complex scholarly acknowledgements and rewards system.
- Taking an active role in promoting active data stewardship will minimize the current perceived risk that journals remain the sole or defining gatekeepers to research outputs, with the reward system for researchers being regulated by publishers alone.
- The Belmont Forum can contribute to addressing the current ineffective data management situation which, at best, leads to research projects delivering unusable data or developing their own data management, exploitation systems and standards (thus increasing the fragmentation). At worst, it leads to making data holdings completely unavailable. All of these current practices seriously limit the impact of the original investment.

Action Theme 2: Data Planning
Near-Term Actions (start within 0-2 years)

1. Belmont Forum members to **appoint a Data e-Infrastructure Officer (DIO) as part of the Coordination Office to oversee Data Plan activities** and to liaise with the Belmont Forum Principals and appropriate members of the scientific community.
2. Belmont Forum to **adopt a common minimum Enhanced Data Plan (EDP)** template, developed in conjunction with all Belmont Forum agencies.
3. DIO to **establish a Security Advisory Board (SAB)** to oversee relevant security issues and contribute to the development of the EDP template.
4. DIO to establish **a Data Policy Advisory Board (DPAB)** to oversee relevant legal issues and contribute toward the development of the Enhanced Data Plan template.
5. DIO to **feed inputs from the SAB and DPAB** to the EDP template, guiding its evolution and development.
6. Belmont Forum members to work with the Data e-Infrastructure Officer to **establish mechanisms for regular review and monitoring of EDP** accomplishments and effectiveness**.**
7. Belmont Forum members to **assign budget allocation for EDP implementation**, including data curation and publishing costs, as part of their standard grant policy.
8. Belmont Forum members to **extend the current reward system** associated with grant awards to include a conditional sharing and sustainable archiving requirement to complete the data lifecycle and to reinforce successful implementation of EDPs.

Medium-Term Actions (start within 1-5 years)

1. Develop and maintain a set of **metrics to assess compliance** with the Belmont Forum-funded EDPs.
2. Belmont Forum members to **review and monitor implementation of EDPs**, starting with funded research projects under the Future Earth initiative.
3. DIO to **identify key repositories** - preferably certified, *trusted repositories* - relevant to the Belmont Forum research agenda to serve as exemplars for the EDPs.
4. DIO to initiate a **gap analysis** to identify data from Belmont Forum-funded research not stored in trusted repositories and work with Belmont Forum agencies and other stakeholders, including the research community, to explore ways of filling gaps.
5. DIO to support the development of a Belmont Forum **research register of software, workflows, policies and standards** (including *vocabularies*, *ontologies*, data models, data structures and computational interfaces) for improved guidance within the EDPs.
6. DIO to **improve the EDP template over time.**
7. DIO to **maintain EDPs in a freely accessible, searchable resource.**

Action Theme 3: e-Infrastructure

Determine international and community best practice in order to inform e-infrastructure policy for all Belmont Forum research

There is a need to develop a roadmap for sharing, shaping and using collaborative data-intensive e-infrastructures. This action theme calls for the Belmont Forum to implement iterative cycles of scoping workshops and international calls for research-driven case studies of the way researchers are using data and e-infrastructures in large interdisciplinary investigations to determine best practice. The workshops will identify the most critical issues, including interoperability at system and service levels, for a variety of cross-disciplinary research applications and communities, and will define calls for case studies. Reports from the case studies will be used to inform the development of Enhanced Data Plans to illustrate weaknesses in existing e-infrastructures and so provide evidence to advise on the technical implementation of future e-infrastructures. This Action Theme will start by using existing cross-disciplinary research projects already being carried out under the Belmont Forum, Future Earth, GEO or other international initiatives, and build on them.

Impact: Examples of good practice in data and information sharing, and in developing and using e-infrastructure, are relatively new, uncoordinated and only rarely have long-term support. Important science,

societal and economic questions require changing methods and practices to facilitate cross-disciplinary research conducted around real world problems rather than in silos. Modern ways of carrying out research allow many domains of environmental and social sciences to come together in much larger and more diverse research teams than in earlier endeavors. New global change research problems are at the international forefront of research and often push the boundaries of the use of data and e-infrastructure.

Action Theme 3: e-Infrastructure
Near-Term Actions (start within 0-2 years)

1. Coordination Office to **appoint an e-Infrastructure Champion** to oversee activities and liaise with the Coordination Office.
2. e-Infrastructure Champion to **convene a series of scoping workshops** to inform the development of an Evaluation Matrix, which will be used to analyze, score, and identify cross-disciplinary case studies, and the broader scope of the relationship between environmental science, data, and e-infrastructure. The workshops should include a mix of environmental, social and health research scientists, data scientists, data-aware engineers, critical stakeholders, international policy makers, decision-makers and advisory bodies.
3. e-Infrastructure Champion to convene a series of scoping workshops to **analyze e-infrastructure applications of existing Belmont Forum-funded projects and other international initiatives,** using the derived Evaluation Matrix to identify critical gaps and barriers, and define any required funding calls and priorities for case studies.

Medium-Term Actions (start within 1-5 years)

1. Belmont Forum to **hold a three-year competitive funding call for case studies** to establish good practice in scenarios prioritized using the Evaluation Matrix.
2. e-Infrastructure Champion to convene additional scoping workshops to **review and analyze reports from case studies to identify strategies to implement best e-infrastructure practices**. Analyses will be used to inform the evolution of e-infrastructure and national investments, and will be reflected in subsequent iterative calls where further strategies may be identified or existing strategies may be enhanced.
3. Belmont Forum members to **define current best practice policy to implement the strategies determined from the analyses** for subsequent cycles of scoping workshops and calls for case studies.

Action Theme 4: Human Dimensions

Support the development of a cross-disciplinary training curriculum to build capability

The objective of this action is to develop a curriculum that can be replicated to: 1) ensure Belmont Forum researchers incorporate data plans into their work; and 2) attract and prepare a new wave of researchers to conduct data-intensive environmental change research, and facilitate their ability to develop new opportunities for cross-disciplinary research through collaboration with existing best practice in data management and data-intensive research.

Researchers are fully stretched just trying to address challenges in their own field. They are naturally focused on their own scientific problems and are competitive with colleagues as well as collaborative. Researchers are therefore reluctant to be diverted into understanding new methods and research practices, particularly if they are unsure of their benefits. This Action Theme addresses a major gap identified in the Open Data Survey questionnaire.

This action theme would: support the development of a holistic training and education curriculum in data-intensive environmental science, for delivery to environmental, social and computer scientists; and launch the curriculum through the creation of a number of international short courses and immersive winter/summer schools. Whilst there is existing training in a number of areas and domains, it is frequently too restricted in its scope and there is considerable evidence of the shortage of skilled people worldwide who have both a well-developed understanding of environmental, social, and health science and skills and knowledge in data science, data-handling and computational methods and technologies. This curriculum will provide a focus and toolkits for training and knowledge dissemination from international and national experts to those who can further apply the knowledge and experience locally.

Impact: National initiatives cannot reach a large enough group of scientists, nor can they involve a wide enough population of appropriate international subject matter leaders. Only a coordinated action of multi-national funding agencies can create the critical mass of competent scientists and ensure the initiative can have a significant and measurable impact. The Belmont Forum can target a large international and interdisciplinary pool of leading scientists and educators, thus obtaining an impact that no single national funding agency can achieve. Cascading the training from Belmont Forum courses nationally will have long lasting impacts that will influence a much wider population. This also has the benefit of allowing further cross communication among disciplinary areas.

Action Theme 4: Human Dimensions
Near-Term Actions (start within 0-2 years)

1. Coordination Office to **appoint a Human Dimensions Champion** to oversee activities, coordinate with activities to implement the other Action Themes, and liaise with Belmont Forum funded projects and Belmont Forum Secretariat. The Human Dimensions Champion should work closely with the Coordination Office to identify existing training initiatives being carried out by Belmont Forum members and others around the globe which address relevant issues.
2. Coordination Office to **create and maintain a database, accessible via a website, of Belmont Forum member and other training initiatives in this area** to ensure that those which are already accessible to international students are not duplicated and are promoted more widely.
3. The Human Dimensions Champion and the Coordination Office to **organize a scoping workshop to design the overall** curriculum for a program of short courses as part of what should become a larger global "virtual university of e-infrastructure" which will deliver skills and knowledge in 'Informatics for Human and Environmental Science'. The Scoping Workshop should address theory and methods for responsibly producing, handling and analyzing environmental, social and health sciences data through local and cloud-based computing tools, data standards and data management best practices for reproducible sciences.
4. Belmont Forum to **initiate a competitive funding call** for the delivery of training courses against this curriculum based on existing successful exemplars and on newly proposed courses.
5. An additional **competitive funding call** should be initiated for the design and delivery of short courses which incorporate the recommendations of this report, including both environmental and relevant aspects of social sciences, as well as informatics. The purpose of the call would be to make nationally-funded short courses Belmont Forum compliant, and then allow people in other Belmont Forum member countries to attend those courses if they are qualified and would benefit from the training.

Medium-Term Actions (start within 1-5 years)

1. Belmont Forum to initiate a subsequent **competitive funding call to develop online training programs**, building on the existing exemplars and results from previous funding calls, and incorporating ongoing developments in data plans and e-infrastructure, to ensure that trainees have the capabilities and tools to pass on their skills to their local communities. These programs should also be aimed at data stewards in order to share best practice internationally.

   Belmont Forum funding should be provided for the continuing provision and updating of the courses, particularly to utilize new and innovative delivery mechanisms. Individual Belmont Forum member agencies may separately allocate funds for appropriate students and scientists to attend international short courses and summer schools to prepare them better to compete in data-enabled activities.

BROADER IMPACTS

Benefits of Acting

This proposed set of initiatives will better enable the Belmont Forum to fulfill its charge to "*to deliver knowledge needed for action to avoid and adapt to detrimental environmental change including extreme hazardous events*". In addition, through internal adoption by individual Belmont Forum members, these recommendations will have much broader impacts for disciplines and programs outside of environmental change research and for organizations engaged in scientific and technical research and operations worldwide.

> ***"Too often, development efforts have been hampered by a lack of the most basic data about the social and economic circumstances in which people live… We must also take advantage of new technologies and access to open data for all people."***
> *- Bali Communiqué of the High-Level Panel, March 2013*

Accelerate the Pace of Scientific Discovery

The recommendations have the potential to transform the way research is conducted by accelerating discovery, increasing the value of research decision-making, and catalyzing changes throughout the economy and society that are of value to all citizens. New scientific discoveries and socio-economic innovation will emerge from tackling the large increase in diversity, volume and rate of growth of multidisciplinary data. Establishing and enabling a cross-disciplinary framework and data-intensive e-infrastructure, with network and computational elements, will allow scientific knowledge to transcend disciplines and address new environmental change problems. Acting now, at a stage early in the development of distributed network solutions and similar

elements of e-infrastructure, means that the Belmont Forum can have extraordinary influence on those specialized developments.

Broaden Dissemination of Best Practice

Actions to adopt and use best practices for research data and e-infrastructure planning and development will ultimately benefit current and future Belmont Forum-funded research, and the general research landscape. This could foster greater trust in research outputs, because data are available for validation and reuse.

Enhance Coordination

Developing coordinated and interoperable data and e-infrastructure includes mapping relevant activities in and among organizations. Mapping will enhance collaboration and general practice within the Belmont Forum, across activities within member agencies and countries, and in institutions involved in the global coordination of environmental and social science information. It will harmonize efforts and organizations, lessen volunteer fatigue, reduce redundancy and duplication of effort, and increase the impact of funding initiatives.

Build Capability

Facilitating international, cross-disciplinary training will increase the potential for broader, global participation in research, and expand human capability and competitiveness. This will result in products and publications of greater benefit to the international community. Students and researchers, especially from developing nations, will also benefit from the opportunity to present their research problems and materials, compare best practice, and network with contemporaries in other countries and disciplines. In itself, this will be an important legacy of the investments described here. Taking all these investments together, they will be transformative.

Ice shelf

The Consequences of Not Acting

Impaired Ability to Respond to Detrimental Effects of Environmental Change

Global change research is extremely time-critical. Given the immediate and long-term risks of environmental change, together with the ever-increasing amounts of research data being generated, much damage would be done to the field of study (Earth) and our ability to start formulating meaningful evidence-driven actions if delays force us to start again or backtrack. Not acting may limit our options and ability to respond to crises, since avoidable errors in decisions occur daily. Decision makers may not know about reasonable options for

adaptation and mitigation because data and knowledge were not shared, or Earth system models will incorrectly assess impacts because they did not incorporate realistic or current data. We can also lose visibility of existing data if they are not curated and made accessible to modern e-infrastructures. Avoiding such errors and loss of data by promoting better access, preservation and use of existing data would yield significant financial savings, reduce distress and save lives.

Lost Opportunities and Squandered Valuable Resources

Not acting will create lost opportunities, delays in achieving Belmont Forum goals, squandering of valuable resources in the form of increased costs to retrofit incompatible data, software and scientific results, and losing data irretrievably. Not acting could also result in losing momentum in the application of globally integrated e-infrastructure for research, which has potentially profound economic and societal consequences. Not acting also means that, in the void of truly globally accepted agreements, special interest developers may be the only option and may drive solutions that are incompatible with environmental and social science research needs.

NEXT STEPS

The Secretariat will widely disseminate this report to foster dialog in the scientific and technical communities on its findings and recommendations. The Belmont Forum organizations will be holding their annual meeting in Oslo, in October 2015. The actions and recommendations made in this report will be on the agenda. In preparation for that meeting, the Group of Program Coordinators from the different member organizations have been asked to consider the activities that most interest their organization and to consider how they might participate going forward. This also provides an opportunity for Belmont Forum members who have not participated in this CRA to register their interest to participate in forthcoming activities. Some organizations external to the Belmont Forum, national and international, have already expressed interest in participating or collaborating with the implementation of the recommendations. Others who wish to do so can contact their national representative in the Belmont Forum or the CRA Secretariat. The Secretariat can also provide briefings to interested parties.

ACKNOWLEDGEMENTS

The project Secretariat gratefully acknowledges the contributions of the Steering Committee in the organization and conclusions of this report, with invaluable guidance and insights from the Group of Program Coordinators and the project Assembly. Supporting documents and evidence can be found on the project *Knowledge Hub* at www.bfe-inf.org. This CRA was supported by the US National Science Foundation (NSF) under Grant No. 1358690 and the UK Natural Environment Research Council (NERC) under Grant Reference NE/LO14391/1. Members of the Steering Committee and Assembly were supported by their individual Belmont Forum participating agencies. Opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF or NERC.

ACRONYMS AND GLOSSARY

Acronyms

CCCO - Communication, Collaboration and Coordination Officer

CRA - Collaborative Research Action

CSIP - Community Strategy and Implementation Plan

DIO - Data e-Infrastructure Officer

DPAB - Data Policy Advisory Board

EDP - Enhanced Data Plan

SAB - Security Advisory Board

SCN - Strategic Coordination Network

Glossary

The following terms have been used within this report and are explained here in more detail. The first instance of the term in the text has been italicized.

Cloud computing

Cloud computing is an on-demand service offering a large pool of usable and accessible virtualized resources, such as hardware (storage, computing, networking), platforms (development and production) and system of software stack (data management, submission, etc.) in infrastructures that provide those services. The hardware and software is what is called Cloud. Public clouds provide those services to external users, usually in a pay-as-you-go manner. Public research clouds are emerging in the landscape today as an alternative to commercial cloud providers.

Data-intensive science

Data-intensive science uses advanced mathematical and statistical methods, computing, and information and communication capabilities to help researchers explore and manipulate massive data sets and enable new insights, not previously possible. In consequence, the speed at which a scientific discipline advances will depend on how well its researchers collaborate with one another and with technologists in areas such as databases, workflow management, visualization, and hybrid and cloud computing technologies.

Data lifecycle

Data often have a longer lifespan than the research project that creates them. Researchers may continue to work on data after funding has ceased, follow-up projects may analyze or add to the data, and data may be reused by other researchers. Well organized, well documented, preserved and shared data are invaluable to advance scientific inquiry and to increase opportunities for learning and innovation. The data lifecycle includes creating data, processing them, analyzing them, preserving them, giving access to them, and then reusing them.

Data literacy

This is the ability to read, create and communicate data as information.

e-Infrastructure

e-Infrastructure is a combination and interworking of distributed hardware resources (storage systems, computing, networking), Information and Communication Technology (ICT) architectures (e.g. Cloud and virtualization, service-oriented architectures are among the most promising architectures today) and digitally-based technology stack (software, middleware), services and tools (data management and access services, High Performance and commodity computing services, virtualization services, networking services, security services), together with the people and organizational structures and policies (access, authorization) needed to support it all.

e-infrastructures enables public and private resource infrastructures with ICT-enabled methods to achieve new, better, faster, and more efficient research, innovation, decision making. There is a growing trend away from delivery of e-Infrastructure as a technology or product in favor of delivery as a service.

Data-intensive e-infrastructure describes a coherent model for e-infrastructure as a service driven by new data-intensive science that makes data an active component of the e-infrastructure, and requires close coordination and interoperability between key services that were historically divided into components: data management, HPC and commodity computing, networking. Data-intensive e-infrastructure is based on access to and analysis of large amounts and of large diversity of new and existing multidisciplinary data in innovative combinations. Long-data storage, curation, inter-operability and certification are just the tip of the iceberg.

Data-intensive e-infrastructures can be described as a new scientific instrument, sometimes referred as a Data-Scope. They are specially designed to provide a seamlessly interacting service allowing users to fluently "observe", analyze and exploit large amounts of data (created not only by scientific instruments and computers but also by processing and collating archived data) from various disciplines such as environment, social, health sciences for fundamental research and society's urgent research applications.

The e-infrastructure data layer is specifically the data element of any e-infrastructure, such as the functions to prepare and preserve data.

High Performance Computing (HPC)

High Performance Computing refers to the use of high-end parallel computing architectures combining capability computing performance (serving a coarse number of specialized computing applications requiring extremely powerful parallel execution, during which computing, communication, and data throughput tasks are tightly coupled) and capacity computing performance (serving an extremely large number of concurrent parallel tasks on a large-scale computing architecture). HPC technology is rapidly changing toward massively parallel multicore and hybrid architectures to deliver extreme scale performance (in computing, data storage and management systems) which requires a change of paradigm in terms of methods and algorithms.

Today, HPC is not only measured by the number of floating-point operations per seconds (flops) at the scale of the petaflop (10^{15} flops) and soon the exaflop (10^{18} flops) but also in terms of data-handling capacity (at the petabyte scale and soon the zetabyte scale) and capability (high performance parallel data management systems). HPC infrastructures and services are provided in a relatively centralized manner from a limited number of large-scale installations (HPC centers). Those infrastructures have diverse geographical constraints, networking environments, organizational models and authorization policies which vary between countries.

Interoperability

Interoperability is the ability of a computer system or software to work with other systems or products without special effort on the part of a user.

Knowledge Hub

The Knowledge Hub is the legacy repository for documents produced during the term of this CRA.

Metadata

Metadata describes other data. It provides information about a certain item's content. For example, a text document's metadata may contain information about how long the document is, who the author is, when the document was written and a short summary of the document. A metadata on a data set will include information about the source of the data, and when and where it was captured or generated, while the data's metadata will describe the content of the data fields.

Open Data

According to the Open Data Handbook (*opendatahandbook.org/guide/en/what-is-open-data/*), "*Open Data is data that can be freely used, reused and redistributed by anyone - subject only, at most, to the requirement to attribute and share alike*". Open Data must be available as a whole, and at no more than a reasonable reproduction cost, preferably by downloading over the internet. The data must also be available in a convenient and modifiable form. Open Data must be provided under terms that permit reuse and redistribution including the intermixing with other data sets. Everyone must be able to use, reuse and redistribute (there should be no discrimination against fields of endeavor or against persons or groups). Open Data enables interoperability: the ability of diverse systems (and organizations) to work together.

Open Science

Open Science is the movement to make scientific research, data and dissemination accessible to all levels of society. Much of the work of science depends on having appropriate tools available to analyze experimental data and to interact with models. Powerful computers are now cheap enough such that significant processing power is within reach of many people. What is often missing is software that lets a researcher choose between models and make sense of the observations, and to test reproducibly the observations and models of other researchers.

Ontology

In information technology, an ontology is a set of concepts - such as things, events, and relations - that are specified in some way (such as specific natural language) in order to create an agreed-upon vocabulary for exchanging information.

Provenance

Provenance of data refers to the process of tracing and recording the origins of data and their movement.

Trusted Repositories and Trusted Data

Trusted Repositories have a mission to provide reliable, long-term access to managed digital resources both now and in the future. Trust is important in two important areas:

- Access to data which may be sensitive in some way or another
- Preservation services which allow the long-term curation and reuse of any data, despite their potentially disclosive nature

In the first case, data owners have to trust a repository not only to provide access solely to authorized users, but also to carry out services such as ingest processing (which includes ensuring the data are appropriately anonymized and internally consistent) and data archiving (managing the data within a secure environment) without disclosing any sensitive information. Trust ought to be transitive: data subjects who trust data owners to look after information about them appropriately should implicitly trust the data archives and repositories who become custodians of these data.

In the second case, data users have to trust the data held in the archive are the same data that have been deposited by the data owners. The data should remain so as repositories migrate to new standards and formats to support long-term preservation. Data provided should be not only usable but also authentic and reliable versions of the data. Data users also have the right to know whether the reproducibility of results will be affected by changes to the data. Increasingly, and especially with international access to data, repositories which have been licensed to provide access to data and are able to assign access to these data to other archives have to trust each other. Data creators, repositories and users also increasingly rely upon each other for services as well as data.

Data has to be stored and distributed together with meta-information (including provenance) for long periods, and data integrity must be proven.

Trusted Data are drawn from carefully selected sources, transformed in accordance with the data's intended use, and delivered in formats and time frames that are appropriate to specific consumers of reports and other manifestations of that data. Trusted Data should have the following six properties: be complete, be current, be consistent, be clean, be compliant and be collaborative. There must be a complete description of the data available and their provenance.