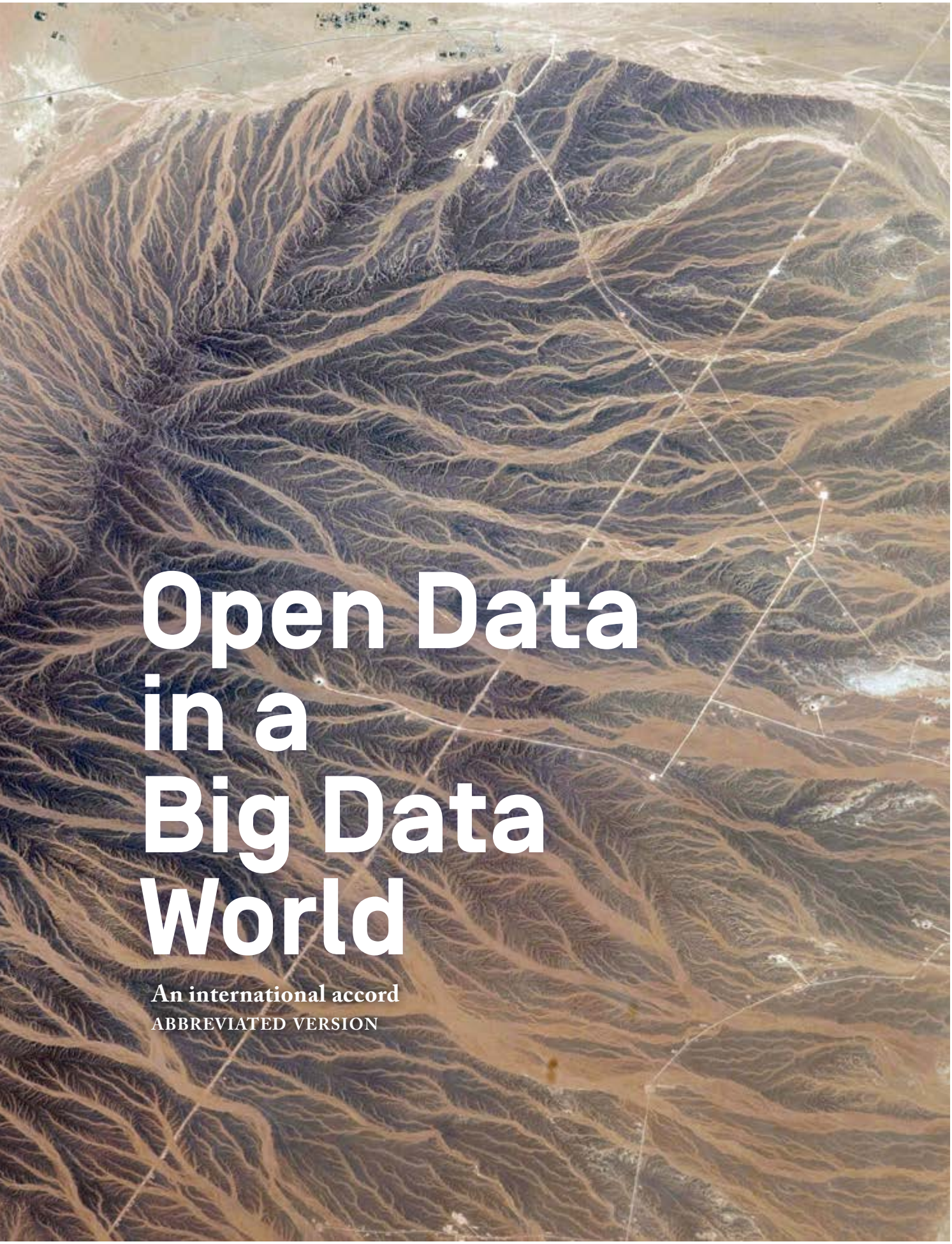


Open Data in a Big Data World

An international accord

ABBREVIATED VERSION

ICSU International Council for Science
iap SCIENCE RESEARCH HEALTH the interacademy partnership
ISSC international social science council
twas

This accord is presented as an outcome of "Science International 2015", the first of a series of annual meetings of four top-level representatives of international science (the International Council for Science – ICSU, the InterAcademy Partnership – IAP, The World Academy of Sciences – TWAS and the International Social Science Council – ISSC) that are designed to represent the global scientific community in the international policy for science arena.

The accord identifies the opportunities and challenges of the data revolution as today's predominant issue for global science policy. It proposes fundamental principles that should be adopted in responding to them. It adds the distinctive voice of the scientific community to those of governments and inter-governmental bodies that have made the case for open data as a fundamental pre-requisite in maintaining the rigour of scientific inquiry and maximising public benefit from the data revolution in both developed and developing countries.

Science International partners will promote discussion and adoption of these principles and their endorsement by their respective members and by other representative bodies of science at national and international levels.

An extended version of this accord is available that includes a deeper rationale for the statements in this document and greater detail about practices designed to support development of open data systems and approaches.

1. The Big Data World

The digital revolution of recent decades is a world historical event as deep and more pervasive than the introduction of the printing press. It has created an unprecedented explosion in the capacity to acquire, store, manipulate and instantaneously transmit vast and complex data volumes, with profound implications for science[1]. The rate of change is formidable. In 2003 scientists declared the mapping of the human genome complete. It took over 10 years and cost $1billion – today it takes mere days and a small fraction of the cost ($1000). "Big data", in which unprecedented fluxes of data stream in and out of computational systems, and "Broad Data" in which numerous datasets can be semantically linked to create deeper meaning, are the engines of this revolution, offering novel opportunities to natural, social and human sciences.

2. The Opportunities

The scientific opportunities of this data-rich world lie in discovering patterns that have hitherto been beyond our reach; in linking and correlating different aspects of systems better to understand their behaviour; in characterising complexity; and in iterating between descriptions of the state of a complex system and simulations that forecast its dynamic behaviour. There are many areas of research where such capacities are deeply relevant: in weather and climate forecasting; in understanding the workings of the brain; in the behaviour of the global economy; in evaluating agricultural productivity; in demographic forecasts; in unravelling histories; and in many of contemporary global challenges such as those of environmental change, infectious disease and mass migration that require combined insights and data from many disciplines.

3. The Challenges

Grasping these opportunities poses serious challenges to the way science is done and organised. Open data are the common, enabling threads.

The Open Data Imperative

The fundamental role of publicly funded research is to add to the stock of knowledge and understanding that are essential to human judgements, innovation and social and personal wellbeing. The technologies and processes of the digital revolution provide a powerful medium through which scientific productivity and creativity can be enhanced by permitting data and ideas to flow openly, rapidly and pervasively through the networked interaction of many minds. If this social revolution in science is to be realised it is vital that we adopt a default position that publicly funded data should be made publicly accessible and re-usable when a research project through which the data have been collected is completed.

Maintaining self-correction

Openness of the evidence (the data) for scientific claims is the bedrock of scientific progress. It permits the logic of an argument to be scrutinised and the reproducibility of observations or experiments to be tested, thereby supporting or invalidating those claims. When a paper making a scientific claim is published, it is essential that the evidentiary data, the related metadata that permit their re-analysis, and the codes used in computer manipulation are made concurrently open to scrutiny to ensure that the vital process of self-correction is maintained. Recent demonstrations in several disciplines of high rates of non-reproducibility of results of published papers emphasise the crucial need to re-invigorate open data processes for a big data world. Openness is not however enough. Data must be intelligently open, meaning that they should be: discoverable, accessible, intelligible, assessable and (re-)usable.

Adapting scientific reasoning

Many of the complex relationships that we now seek to capture through big- or broad-, linked data lie far beyond the analytical power of many classical statistical methods. They require deeper mathematical approaches including topological methods to ensure that inferences drawn from big data and broad data are valid. Data-intensive machine-analysis and machine-learning are becoming ubiquitous, and have major implications for scientific discovery. The complexity of patterns that machines are able to identify are not easily grasped by human cognitive processes, posing profound issues about the human-machine interface and what it might mean to be a researcher in the 21st century.

Ethical constraints

The open data principle has ethical implications for researchers and research subjects. It can appear to override the individual interests of the researchers who generate the data, such that novel ways of recognising and rewarding their contribution need to be developed. The privacy of data subjects needs to be protected. In a regime of open sharing in which data are passed on from their originators, there is loss of control over future usage, whilst anonymisation procedures have been demonstrated to be unable to guarantee the security of personal records.

Open global participation

Big data and open data have great potential to benefit less affluent countries, and especially least developed countries (LDCs). However, LDCs typically have poorly resourced national research systems. If they cannot participate in research based on big and open data, the gap could grow exponentially in coming years. They will be unable to collect, store and share data, unable to participate in the global research enterprise, unable to contribute as full partners to global efforts on climate change, health care, and resource protection, and unable fully to benefit from such efforts, where global solutions will only be achieved if there is global participation. Thus, both emerging and developed nations have a clear, direct interest in helping to fully mobilize LDC science potential and thereby to contribute to achievement of the UN Sustainable Development Goals.

1 The word "science" is used to mean the systematic organisation of knowledge that can be rationally explained and reliably applied. It is used, as in most languages other than English, to include all domains, including humanities and social sciences as well as the STEM (science, technology, engineering, medicine) disciplines.

Seizing the opportunity

Effective open data can only be realised if there is systemic action at personal, disciplinary, national and international levels. Although science is an international enterprise, it is done within distinctive national systems of responsibility, organisation and management, all of which need to respond to the opportunity. Research funders and research performing institutions should fund and implement processes that lighten the burden on researchers of making data intelligently open and that support open data processes.

Increasing numbers of research communities have discovered the benefits of sharing data, in fields as varied as linguistics, bio-informatics and chemical crystallography, and have made major strides in realising benefit for their disciplines through international collaboration in facilitating access and use of open data.

Responsibilities also fall on international bodies, such as the International Council for Science's (ICSU) Committee on Data for Science and Technology (CODATA), its World Data System (WDS) and the Research Data Alliance (RDA), to promote and support developments of the systems and procedures that will ensure international data access, interoperability and sustainability.

Open science and public knowledge

The idea of "open science" has developed in recognition of the need for stronger dialogue and engagement by the scientific community with wider society in addressing many current problems through reciprocal framing of the issues and the collaborative design, execution and application of research. There are, of course, legitimate limits to openness, such as the need to protect security, privacy and proprietary concerns through judiciously applied mechanisms. There are also countervailing trends towards privatisation of knowledge that are at odds with the ethos of scientific inquiry and the basic need of humanity to use ideas freely. If the scientific enterprise is not to founder under such pressures, an assertive commitment to principles of open data, open information and open knowledge is required from the global scientific community.

4. Principles of Open Data

Such is the importance and magnitude of the challenges to the practice of science from the data revolution that Science International believes it appropriate to promote the following statement of principles of open data.

Responsibilities

Scientists

i. Publicly funded scientists have a responsibility to contribute to the public good through the creation and communication of new knowledge, of which associated data are intrinsic parts. They should make such data openly available to others as soon as possible after their production in ways that permit them to be re-used and re-purposed.

ii. The data that provide evidence for published scientific claims should be made concurrently and publicly available in an intelligently open form[2]. This should permit the logic of the link between data and claim to be rigorously scrutinised and the validity of the data to be tested by replication of experiments or observations. To the extent possible, data should be deposited in well-managed and trusted repositories with low access barriers.

iii. Research institutions and universities

have a responsibility to create a supportive environment for open data. This includes the provision of training in data management, preservation and analysis and of relevant technical support, including library and data management services. Institutions that employ scientists and bodies that fund them should develop incentives and criteria for career advancement for those involved in open data processes. Consensus on such criteria is necessary nationally, and ideally internationally, to facilitate desirable patterns of researcher mobility. In the current spirit of internationalisation, universities and other science institutions in developed countries should collaborate with their counterparts in developing countries to mobilise data-intensive capacities.

iv. Publishers

have a responsibility to make data available to reviewers during the review process, to require intelligently open access to the data concurrently with the publication which uses them, and to require the full referencing and citation of these data. Publishers also have a responsibility to make the scientific record available for subsequent analysis through the open provision of metadata and open access for text and data mining.

v. Funding agencies

should regard the costs of open data processes in a research project to be an intrinsic part of the cost of doing the research, and should provide adequate resources and policies for long-term sustainability of infrastructure and repositories. Assessment of research impact, particularly any involving citation metrics, should take due account of the contribution of data creators.

vi. Professional associations, scholarly societies and academies

should develop guidelines and policies for open data and promote the opportunities they offer in ways that reflect the epistemic norms and practices of their members.

vii. Libraries, archives and repositories

have a responsibility for the development and provision of services and technical standards for data to ensure that data are available to those who wish to use them and that data are accessible over the long term.

Boundaries of openness

viii. Open data should be the default position for publicly funded science. Exceptions should be limited to issues of privacy, safety, security and to commercial use in the public interest. Proposed exceptions should be justified on a case-by-case basis and not as blanket exclusions.

Enabling practices

ix. Citation and provenance

When, in scholarly publications, researchers use data created by others, those data should be cited with reference to their originator, to their provenance and to a permanent digital identifier.

x. Interoperability

Both research data, and the metadata which allows them to be assessed and reused, should be interoperable to the greatest degree possible.

xi. Non-restrictive reuse

If research data are not already in the public domain, they should be labelled as reusable by means of a rights waiver or non-restrictive licence that makes it clear that the data may be re-used with no more arduous requirement than that of acknowledging the producer.

xii. Linkability

Open data should, as often as possible, be linked with other data based on their content and context in order to maximise their semantic value.

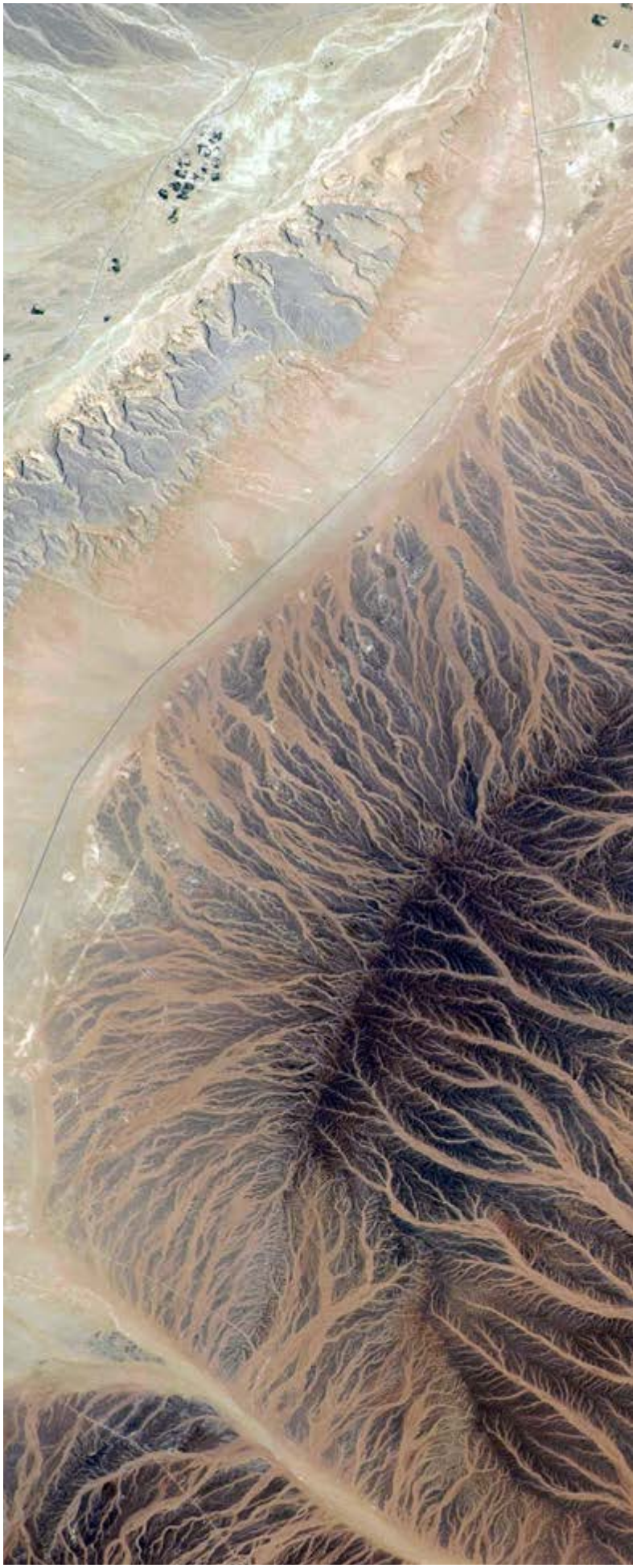
2 Refer to the full text document at http://www.science-international.org

This document was prepared by an ICSU-IAP-ISSC-TWAS working group of:

- **Geoffrey Boulton,**
  University of Edinburgh and President of CODATA,
  Working Group Chair
- **Dominique Babini,**
  University of Buenos Aires and CLACSO (ISSC representative)
- **Simon Hodson,**
  Executive Director CODATA (ICSU representative)
- **Jianhui Li,**
  Chinese Academy of Sciences, CNIC (IAP representative)
- **Tshilidzi Marwala,**
  University of Johannesburg (TWAS representative)
- **Maria G. N. Musoke,**
  Makerere University, Uganda (IAP representative)
- **Paul F. Uhlir,**
  Scholar, US National Academy of Sciences (IAP representative);
  Independent Consultant, Data Policy and Management
- **Sally Wyatt,**
  Maastricht University, & eHumanities, KNAW (ISSC representative)

The extended version of this Accord is available at
http://www.science-international.org

Hard copies are available from
International Council for Science (ICSU),
5 rue Auguste Vacquerie, 75116 Paris, France.



ICSU
International Council for Science

iap SCIENCE RESEARCH HEALTH
the interacademy partnership

ISSC
international social science council

twas

www.icsu.org
www.interacademies.net
www.worldsocialscience.org
www.twas.org