Exploiting the digital revolution: developing capacity and integrating data across the disciplines of science

Executive summary

The digital revolution is of world historical significance. It enables powerful and unprecedented ways of analysing and understanding the complex phenomena that are vital parts of the human condition and the planetary function. This revolution has created two major opportunities to advance the disciplines of scientific inquiry:

- *by exploiting emerging data collections to unleash the potential for game-changing discoveries in all scientific disciplines;*
- *by integrating these data across diverse research domains to create interdisciplinary knowledge and thereby the capacity to address the many inherently complex, global problems that challenge modern science.*

Addressing these challenges depends fundamentally on foundational work in developing shared vocabularies and organised knowledge systems that both permit the discovery and use of data specific to disciplinary priorities and to ensure the interoperability required for the integration of data from diverse disciplines. Pro-active steps are needed to create the necessary cross-discipline scientific data communities.

In June 2017, the International Council for Science (ICSU) and its Committee on Data for Science and Technology (CODATA) brought together international scientific unions and associations of ICSU and the International Social Science Council (ISSC) that have made major strides in this area of work, as well as other organisations that curate standards and vocabularies for particular disciplines. The objective of the meeting was to develop an action plan to realise the full potential of the data science, technologies, and infrastructures currently being created by specific disciplinary groups and expand those efforts on an inter- and trans-disciplinary basis.

The meeting identified key opportunities of the digital revolution and how they can be achieved. Priorities for action include:

- *the need for examples of the benefits that have already been realised by specific disciplinary groups and inter- and trans-disciplinary projects;*
- *the need to extend activities to disciplinary fields that have not yet developed strategies , for developing interoperable vocabularies, standards and models, and for the creation of effective "information communities";*
- *there must be a major effort to achieve interoperability within and between disciplines, without this, the national and regional initiatives to create cloud or platform technologies designed to provide services to support data priorities will fall far short of their potential;*
- *international scientific unions and associations, and the international councils of which they are members, are uniquely qualified for this task, and their engagement is essential if its promise is to realised;*

- *there is a need to develop a flagship programme on one or more major global challenge themes to develop, demonstrate and apply the methods of linking and integrating data from across the disciplines in the production and use of actionable knowledge.*

Such a programme will entail a long-term, decadal commitment. It will convene and support the scientific members of ICSU and ISSC, serve as a mechanism for their engagement with relevant international research initiatives, significantly strengthen their data capacities and relate to the priorities of research funding bodies such as the Belmont Forum. The immediate next step is a major ICSU-CODATA workshop in November 2017 to bring together the full range of scientific international unions and associations with organisations working on complex global problems to sharpen the design of the flagship project and create the international, multi-disciplinary data community needed to convert these opportunities into solutions.

Contents

The challenge

1. The means whereby data, information and knowledge are acquired, stored, analysed and communicated are fundamental determinants of human material and social progress. The explosion of digital technologies in the last two decades has dramatically increased the power and efficiency of these processes. They have ushered in a digital revolution that has more profound and pervasive impacts than the invention of the printing press 550 years ago, and with enormous implications for economies, societies and for science.[1]

2. The Earth now is, and the Future Earth increasingly will be, a networked Earth; with individuals, societies and their institutions, and almost all powered devices, generating, receiving and creatively utilising exponentially increasing data and information fluxes. Although the ways that citizens and institutions adapt to and use the capacities of this new Earth are highly uncertain, what is not in doubt is the magnitude of the impact that global networking has already had and the potential that it has for future disruptive change.

3. There has been an explosive growth in the diversity and volume of data available to scientists, from surveys, sensors and simulations, which has created a novel potential for new understanding of complex systems on all scales, from the molecular to the cosmic, and all in areas of human concern, from cultural artefacts to local health systems to global sustainability. Grasping these opportunities depends upon the willingness of the scientific community to grasp the challenge of open data,[2] the ability of scientists to discover relevant data from multiple distributed sites, and the capacity to semantically integrate data from disparate disciplines that bear on complex, cross-disciplinary problems where the use of machine-learning algorithms from artificial intelligence can reveal relationships in multi-dimensional data to create profound understanding of complex systems and processes.

Enhancing capability, grasping opportunity

4. These latter abilities depend fundamentally on some very basic, essential procedures that form the vital foundations for finding the data that researchers may need, gaining access to the data and associated metadata, and combining or integrating them. Success in this depends on the degree to which data have common or translatable vocabularies, follow common structures and include the requisite structured metadata that enable automated use and interpretation. Without such procedures and standards, the disciplines of science will be unable to utilise their data resources outside their immediate domains and will be unable to exploit the rapidly expanding universe of possibilities to best effect.

5. There is however a further horizon. In 2003, Tim Berners-Lee and colleagues proposed that the World Wide Web, which discovers and produces electronic documents on request, could become a "semantic web" that allows data to be shared and reused across applications, enterprises, and community boundaries, and machine-integrated to create knowledge. The most profound potential lies in understanding the behaviour of complex systems, including that most challenging, but vital domain, of the interactions between human and non-human systems.

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

[2] **Open data** are data that are freely available to everyone to use and republish as they wish, without restrictions from copyright, patents, or other mechanisms of control.

6. However, the semantic linking of data so that they can be queried or integrated in ways that reveal highly complex, multi-dimensional patterns is dependent on the development and use of compatible standards for the discovery, access, sharing, use, semantic linking, comprehension, interpretation and retention of these data through the development of an infrastructure of persistent identifiers, vocabularies and ontologies as well as the services needed to make use of them. Without definition of the standards that underlie data sets from contributing scientific disciplines, it will be difficult to programmatically analyse, reuse and integrate data sets generated by different disciplines and from different countries to address the most pressing global scientific challenges that can only be addressed effectively through such integration. There is thus an urgent need to develop coordinated approaches to standards development that will permit the interoperability necessary for efficient data integration and reduce the replication of effort and proliferation of incompatible practices. Without this, many of the complex global challenges that science is called upon to address, including the United Nations Sustainable Development Goals, will not be able to capitalise from the increasing global data resource.

7. These considerations lead to two important conclusions:
   - that social and technical procedures should be adapted as necessary within disciplines to discover, manipulate and share data in creative and mutually advantageous ways;
   - that procedures should be adapted and developed that permit data from diverse sources and different disciplines to be integrated and linked with the minimum effort and maximum benefit to researchers.

Key priorities for 21st century science

8. It is for the reasons described in ¶4-7 that ICSU-CODATA convened an interdisciplinary meeting in Paris on 19-21 June 2017 (Appendix 1) of individuals from groups that have made major strides in developing data resources and services that are highly beneficial to their own research communities, together with a range of data specialists, to discuss:
   - Whether and how it might be possible to stimulate the efforts of those disciplines that have hitherto been less energetic in developing mechanisms to creatively exploit their own data resources. The role of ICSU as a membership body for 31 international science unions as well as 122 national members could be crucial here, whilst the possible merger of ICSU with ISSC offers yet further opportunities.
   - How science might work progressively towards the ambitious objective of integrating and semantically linking data derived from different disciplines that relate to the same or coupled phenomena through the coordination of standards. This capacity would create profound new potential for understanding, and a critical pathway to understanding phenomena of great complexity that underpin the major global challenges to 21st century science.

Evidence of scientific benefit

9. Already international standards bodies such as the International Organisation for Standardisation (ISO), the World-Wide Web Consortium (W3C) and the Open Geospatial Consortium (OGC) have developed technical standards that are applicable to the interchange of scientific data such as the Geography Markup Language (GML), to measurements from observations and sensors, to spatial coordinate systems, and to metadata standards such as the Dublin Core Metadata,[3] which have been ratified as ISO Standard 15836:2009,[4] and to linkage, visualisation and semantic annotation capabilities. What these technical groups will not do, and cannot do, is to develop domain-specific

[3] http://www.dublincore.org/metadata-basics
[4] http://www.iso.org/iso/search.htm?qt=15836&searchSubmit=Search&sort=rel&type=simple&published=on

standards necessary for much of the data collected as part of scientific research. Science conventions such as, for example, ISO standards, should be regarded as the top tier of a hierarchy of data standards, containing a small subset of all the standards/conventions needed within science. Discipline-specific standards, can only be created by disciplinary experts. They fall naturally into the above hierarchy and to be effective should have their scope and identification of their individual properties defined at the outset.

10. There are many disciplines and sub-disciplinary areas that have made great strides in making data resources from their fields openly and widely available. They have developed tools and services that enable data to be creatively used, to demonstrate the scientific value and potential that has been released through these processes and that the value of these changes justifies their cost in time and money. They include major initiatives in astronomy, bio-informatics, crystallography, social sciences, nano-technology, archaeology and the geosciences. Examples of achievements in these areas are summarised in Appendix 2. In several cases the relevant international science unions such as the International Astronomical Union, International Union of Crystallography, International Union of Geodesy and Geophysics and the International Union for Geological Sciences have set up specific groups on data and information. As yet however there has been relatively little coordination between them.

The basis for interoperability: vocabularies, standards, organisations, and the choice and adoption of standards

Data sharing expectations

11. Standards for Interoperability are imperative for the use, exchange and sharing of information to ensure that digital research outputs are FAIR, that is: Findable, Accessible, Interoperable and Reusable.[5]

Vocabularies

12. To be understood, to avoid ambiguity or misinterpretation, each concept within a dataset (data values, column headings, methods, instruments, protocols, classifiers …) must be uniquely defined. Concepts should be named at the aggregation level of usage, so that each *item* from a set or list is accessible individually (e.g. unit-of-measure or classifier), as well as the list-as-a-whole for context. A description of each concept must be available on demand so that data producers and consumers can be confident about their meaning. Concepts come in sets that are conventionally presented in a list or *vocabulary*. Vocabularies can be represented with different levels of expressiveness of formality, some of which are application specific, but some of which are common across domains and disciplines. Each set is typically managed under a common governance arrangement, and has a clearly defined scope.

13. In order for particular sets of names (vocabularies) to be widely used, they must be assigned or curated by a trusted organization. Trust may develop because of association with a recognized authority, or merely through widespread usage and continuous maintenance.[6]

[5] Wilkinson et al., *Nature Scientific Data,* March 2016. DOI: 10.1038/sdata.2016.18
[6] In which case formal 'adoption' by a recognized authority might be considered.

Standards and communities

14. Groups that use common information standards, such as vocabularies, file-formats and exchange protocols can be regarded as an "information community". A '***standard***' is a method or agreement used within that community, with an understanding or expectation that new users are welcome to join the community (if new members are *not* welcome, then it is a contract, not a standard). Standards with the same scope may exist for technical, social or historical reasons, used by different communities (e.g., different lists of units of measure or chemical substances). Information communities come in different sizes. In general we would like science information communities to be as large as possible, so the competence and governance arrangement of the standards that they use has a broad reach.

15. Content standards, file formats and data structures, are crucial to interoperability and in order to make datasets open to transparent interpretation, verification and exchange. The uptake of content standards is vital for high-quality, reproducible research and for the integrative analysis and comparison of heterogeneous data from multiple sources, domains and disciplines. When a content standard is mature and appropriate, standard-compliant software systems and applications become available. These may then be channelled to the appropriate stakeholder community, which in turn can recommend them (in data policies) or use them to facilitate a high-quality data cycle, from data generation to standardization, and through to publication and subsequent sharing and reuse.

Organisations

16. Standardization activities are numerous and diverse, from those driven by large organizations with industrial strength to scientific disciplinary bodies organised internationally, to grass root activities involving a small group of collaborators. Stakeholders participating in these efforts are involved in managing, serving, curating, preserving, publishing or regulating data and/or other digital objects. They are often - but not always - not only producers but also end users of standards. Standards organizations have varying level of formality (e.g., some are legal entities, though the majority are ad hoc working groups), membership types (e.g., open and free vs members only), operational approaches (e.g., organized in formal committee, or as open working groups) and funding levels. The sustainability, authority and governance of organisations that develop, manage, review and adjust standards are important issues in the open science and open data ecosystem, though they are potentially vulnerable to threat.

Choosing and adopting standards

17. Although there is widespread agreement that open, community-developed standards are critical, there is little consensus on which data standards should be used, the criteria by which standard should be chosen, or even what constitutes a data standard. There is a fundamental difference between an ad hoc list of terms, offered as a pragmatic solution to a data terminology issue and a widely adopted and implemented standard with accepted procedures for community input and governance. This points to the need for information resources on vocabularies and standards and a maturity model by which they can be assessed in relation to a range of criteria.

18. For example several thousand different standards exist in the life, environmental and biomedical sciences. In these areas FAIRsharing (formerly BioSharing) is building a comprehensive curated resource that maps this landscape. As an informative resource, FAIRsharing ensures that standards are findable and accessible (similar to the data to which the FAIR principles apply). As an educational resource, FAIRsharing provides the indicators necessary to monitor the development, evolution and integration of standards. By interlinking standards, databases and data policies (from funders, journals and other organizations), FAIRsharing guides users to discover those standards that are implemented by databases and to find the policies that refer to

them, providing evidence of use and other important indicators that users take into consideration when selecting a resource (see Appendix 2). FAIRsharing also crowdsources information to update and curate the description and status of each standard, ranging from 'ready to use' through 'in development', 'uncertain' to 'deprecated' for those standards that are no longer maintained.

19. FAIRsharing is used and adopted by many journals, publishers, research support organizations and research infrastructure programmes. FAIRsharing is a resource of the ELIXIR Interoperability Platform, and operates as a working group under the RDA and the FORCE11 umbrella, reaching out to several disciplines and communities.

20. A report commissioned by the Wellcome Trust contains further information and specific examples of standardization efforts and related challenges and opportunities, especially focussing on the broad life, environmental and biomedical sciences, where a wealth of activities exists.[7] In addition, like any other digital object, standards in general and content standard more specifically have a life cycle. The report of a workshop organized by the National Institutes of Health (NIH) Big Data to Knowledge Initiative (BD2K) on community-driven content standards provides an invaluable insight on different issues pertaining to each phase of the life cycle (i.*e.*, formulation, development and maintenance), showing that communities' social and technical approaches to common problems are diverse.[8]

Extending disciplinary engagement

21. Many disciplines and scientific unions and associations are either not engaged with data standardization or have not developed shared vocabularies and standards. Where shared resources do exist, a "preferred authoritative standard" may not have been clearly endorsed. The expressivity of existing standards is often limited, which impacts on their re-use. . It is important that all relevant areas of research understand the level the level of data standardization that is required in their discipline for effective discovery, access, and use of data, and the problems that need to be solved to use data in multi- and trans-disciplinary modes.

22. The increasing numbers of disciplines, scientific unions and associations that have created "information communities" to engage with the digital challenge have established, often through trial and error, important lessons about what works and what does not, and their experience forms an important guide to later-comers:
    a. Collection of high-quality data is facilitated when there is prior agreement about data collection, data format and metadata standards.
    b. Easily understandable and user-friendly implementation mechanisms are key to the adoption of standards by the research community. Data and metadata standards that are not easily usable tend to be ignored. Web-based tools are ideally suited to this since they require no special software to be installed or learnt, and are inherently distributed.
    c. Traditional QA/QC, statistical and visualization approaches do not typically scale to big, multidisciplinary data. New algorithms and approaches are often required.
    d. Vocabularies and lexicons used in one scientific discipline or domain are not universally understood and significant effort and engagement are necessary to bridge disciplinary boundaries. Sharing across disciplines is however essential, and must be supported either through cross-disciplinary coordination, 'mappings' of concepts from one discipline to another, or by the development of core models that express the common elements of science data.
    e. Big science challenges often demand big data and such data can normally only be acquired and processed using machine-automated approaches. Successful use of big data depends on the

[7] Sansone and Rocca-Serra, 2016. https://doi.org/10.6084/m9.figshare.4055496.v1
[8] https://datascience.nih.gov/sites/default/files/bd2k/docs/ExecSumm_CBDMSworkshopFEB2015.pdf

degree to which data follow common structures and include the requisite structured metadata that can enable automated interpretation of the data.

f. Groups that have had an engagement strategy with their community coupled with the offer of technical enhancements, such as those listed in Appendix 1, have succeeded in attracting widespread use of their services.

The grand technical challenges of the digital revolution

23. The first major challenge lies in extending the capacity for efficient and rigorous use of digital data resources across the whole range of the research enterprise. Naturally enough, those research areas that have made the greatest strides in developing powerful capacities for beneficial exploitation of the potential of the modern data environment are those where data streams can be precisely standardised and codified. It is a more complex task in those research areas where data are highly diverse, where there are different historical traditions of meaning and nomenclature, where data is not "born digital," and where they cannot readily be translated into standard formats. Such cases pose severe problems for semantic linking and integration between the varied data streams that reflect different aspects of the complex, coupled phenomena that lie at the heart of many global challenges. Problems of semantic linking and integration are particularly evident in relation to data from the social sciences and humanities, particularly qualitative data, without which – for example – the human, societal and economic response to global change cannot be understood.

24. The second major challenge lies in the development of processes that will enable semantic linking and integration of datasets from disparate areas of research that bear on the same phenomenon, to greatly enhance the power of inter- and trans-disciplinary research. Even where rigorously defined and widely applied standards exist, the diversity of disciplinary data standards and solutions inhibits this. Coordinated and systematic approaches to the development of data standards with a view to greater interoperability and semantic enrichment is most likely to be achieved through attack on a major, complex research problem, of value in itself but also as a demonstrator of effective approaches to interoperability. Existing data and vocabulary models often fail to capture the specific semantic nuances of complex disciplinary data that extend beyond simple hierarchical relationships. Where existing standards are well entrenched in existing software and databases, adoption/use of a new standard model will be more likely if translators or common interchange formats are developed to allow data to continue to be managed in existing data systems but can easily be exported to new models more suited to integration with other data types.

25. If data is in a precisely codified and standardised form (i.e., born semantic), and if data streams can be integrated as in ¶24, then depending on the richness of the known semantic relationships between multi-dimensional data streams, artificial intelligence machine learning tools can be used to create profound understanding of complex systems and processes. Figure 1 illustrates how the richness of semantic content determines the extent to which algorithms of increasing reasoning power can be used to derive important relationships in complex phenomena.

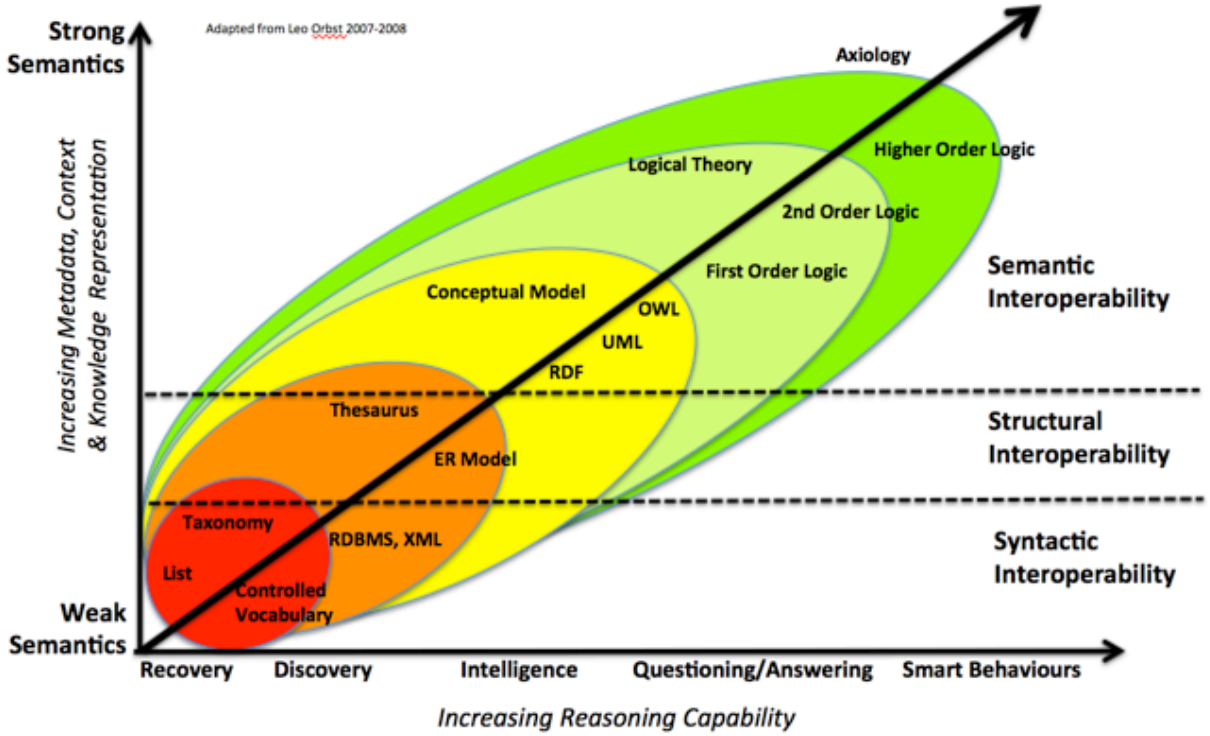

Figure 1. The relationship between semantic richness and reasoning potential.

Extending capacity across the breadth of research

26. Developing the capacity within the scientific enterprise to characterise and understand complexity as described above would be a major contribution to the toolbox of 21st century science. Doing so across the whole range of scholarly inquiry is a daunting task, involving:
    - explaining the potential of vocabularies, standards and semantic linking in manipulating the data streams of the digital world to create benefit to individual research domains;
    - providing technical guidance and ways to access the expertise they need to those unions and associations, in their role as the international representatives of disciplinary research, that wish to develop or define standards or other infrastructures in building a semantic web;
    - providing guidance to unions/associations about exposing vocabularies in controlled name space so that they are persistent, sustainable and governed so that community can control their evolution.

    There are, in parallel, needs for:
    - training, instruments, software, platforms, and databases to support the use of preferred authoritative standards;
    - a go-to place or trusted host for information on preferred discipline standards;
    - information on data approaches for inter- and trans-disciplinary science.

27. It is important to recognise that these priorities need to be addressed along disciplinary axes, and for outcomes to achieve international consensus. They are orthogonal to national and regional initiatives for cloud or platform technologies designed to provide services to support data priorities across the disciplines in their jurisdictions. However, it is important to recognise that without a healthy and pervasive disciplinary axis, the potential of national/regional infrastructure will be under-realised. The former is a fundamental task for which the international scientific unions and associations, and the international councils of which they are members are uniquely qualified.

28. To provide a basis for a coordinated effort to address these priorities, CODATA has created a Task Group on *Coordinating Data Standards amongst Scientific Unions* (http://www.codata.org/task-groups/coordinating-data-standards) with the following objectives:
    - Identify science unions/associations that have a commission on data, or a point of contact for on standards to be developed/governed/endorsed by their unions/associations;
    - Take a leadership role in raising awareness of standards endorsed by and/or being developed by the unions, to assist with authoritative standards and minimise duplication;
    - Create a web page linked to repositories for data models, information standards, vocabularies, ontologies, etc., for each union, avoiding duplication with existing portals (e.g FAIRsharing);
    - Determine a standards "maturity model' adapted from the 5-star model and the AGU Data Maturity Framework (http://5stardata.info/en/) (http://dataservices.agu.org/dmm/) as a guide to users on usability of standards, to developers on maturity of their standards and to assist in ensuring "fitness for purpose"; combined with the '"indicator of status" of FAIRsharing;
    - Provide best practice examples for the development and application of the standards, guidance on governance frameworks, with linkages to the Research Data Alliance, and national efforts such as the Australian National Data Service (ANDS), the Earth Science Information Partners (ESIP), EU 2020 projects, etc.;
    - Provide guidelines to the scientific community on adherence to these standards and promote their benefits for discovery and access to data.

29. Driving this development forward across the breadth of the research community will require both the engagement of bodies that represent that breadth and a long term, decadal effort. The engagement of the International Council for Science and the International Social Science Council, OECD, and possibly UNESCO, with these priorities will be important. Their engagement should be sought because of their convening influence with the international scientific unions and associations that are their members, as custodians of long term scientific priorities and as influential voices in supporting approaches to funders. It is also important that broader collaboration and coordination are sought with the medical, engineering and humanities communities.

30. The process initiated at the June 2017 Paris meeting was designed to look beyond the work of this Task Group and to begin development of a road map for the destinations summarised in paragraphs 23-25. It is CODATA's intention to create a formal Commission to carry this work forward, in consultation with the groups represented in Paris and listed in Appendix 1, and in consultation and collaboration with bodies such as the Research Data Alliance and the World Data System.

A flagship project

31. The priorities outlined in ¶23 relate to the data capacities of individual disciplines, but the most ambitious priority is to realise the potential of inter-disciplinary semantic linking and integration (¶23) as shown in Figure 1. At the same time the many so-called global challenges and the imperatives of the sustainable development goals pose urgent needs for understanding the complex systems of which they are almost invariably parts, and for which linked semantic data is a fundamental key. For this latter reason that we believe a flagship project to address a global challenge and building an "information community" around it is an important priority for:
    - delivering action-oriented knowledge for a key human priority;
    - developing standardised approaches for inter- and trans-disciplinary semantic linking;
    - creating an influential exemplar of procedures and processes for major analogous initiatives by the scientific community.

32. The chosen project should satisfy a number on important criteria, that:
    - there is a strong demand for solutions from the international community, including political support;

- there is a community of practice able to provide specialist domain knowledge that is enthusiastic to collaborate with data experts in creating an information community and undertaking a major project;
- a broad range of disciplinary expertise is involved in work in the domain, and that they include both natural and social science disciplines;
- the necessary FAIR data exists;
- the contributing disciplines have, as far as possible, data, metadata and other interoperability standards, and established information communities;

33. A series of recognised "global challenges" have been identified that could be the focus for a flagship project, and which have the potential to satisfy the above criteria. All depend upon the integration of data from a variety of sources and on standards and vocabularies which are essential to interoperability of data across domains:
    - Clean water/clean air
    - Disaster risk reduction
    - Drug development and antibiotic resistance
    - Data sharing for public health: transmissible diseases
    - Impact and mitigation of sea level rise on coastal cities and ecosystems
    - National and international security in the cyber world
    - Preservation of cultural and natural heritage
    - Sustainable agriculture
    - Deforestation
    - Invasive species
    - Social and economic consequences of ageing populations
    - Maintaining biodiversity
    - Poverty
    - The future of humankind on earth

34. These options are currently being scoped against the criteria in ¶32. All require a wide range of inputs from natural, social, medical and engineering science, and many of them would benefit from input from the humanities.

Next steps

35. The development of the above programmatic concepts and of the data science/domain science and funding communities actions that are need to sustain and support them will require consultation and joint action. The immediate next steps in this process will be as follows:
    1) Review and development of this paper by participants in the Paris meeting.
    2) Discussions with the Executive Directors of ICSU and ISSC about how this concept could best be developed to engage with their international unions and associations (note the possibility that both may merge after October 2017) and how to engage with international funders.
    3) Presentation of this report to the Executive Boards of ICSU and ISSC to seek their views and support for further developments.
    4) Exploratory discussions with key representatives of relevant international research initiatives, e.g. Future Earth, IRDR, WCRP, facilitated by ICSU/ISSC.
    5) Development of the design of a flagship project and the concept of the Commission to be discussed and refined in the November meeting (see 6 below).
    6) Funding has now been obtained for a major workshop to take place in mid-late November 2017 to take these issues further and to involve a larger representation of scientific unions and associations.

Appendix 1. Participants in the June 21-22, 2017 meeting in Paris

1.	Phil	**Archer**	W3C and VRE4EIC Project
2.	Franz-Josef	**Behr**	ICA
3.	Hugo	**Besemer**	FAO and IGAD
4.	Niklas	**Blomberg**	Elixir
5.	Geoffrey	**Boulton**	CODATA
6.	John	**Broome**	CODATA
7.	Simon	**Cox**	CODATA TG and CSIRO
8.	Markus	**Döring**	GBIF, Darwin Core
9.	Rachel	**Drysdale**	Elixir
10.	Patrick	**Garda**	French Ministère de l'Enseignement Supérieur et de la Recherche
11.	Philippe	**Gaucher**	French Ministère de l'Enseignement Supérieur et de la Recherche
12.	Helen	**Glaves**	Oceans Data Interoperability Platform (ODIP); BODC
13.	Heide	**Hackmann**	ICSU
14.	Bob	**Hanisch**	NIST
15.	John	**Helliwell**	IUCr
16.	André	**Heughebaert**	Belgian Biodiverity Platform and GBIF
17.	Simon	**Hodson**	CODATA
18.	Andreas	**Kempf**	ZBW
19.	Dimitris	**Koureas**	RDA Domain Interoperability Framework, TDWG
20.	Claire	**Melamed**	Global Partnership for Sustainable Development Data
21.	Bill	**Michener**	DataONE and Dryad
22.	Andrea	**Perego**	European Commission
23.	Jean-Luc	**Peyron**	IUFRO
24.	François	**Robida**	IUGS
25.	John	**Rumble**	CODATA Nanomaterials WG
26.	Alena	**Rybkina**	IUGG
27.	Susanna	**Sansone**	BioSharing
28.	Ingo	**Simonis**	OGC
29.	Maria	**Uhle**	Belmont Forum
30.	Jean-Pierre	**Vilotte**	IPGP
31.	Joachim	**Wackerow**	GESIS - Leibniz Institute for the Social Sciences
32.	Sally	**Wyatt**	Huygens Institute
33.	Lesley	**Wyborn**	CODATA TG, ANU and AuScope

Appendix 2. Examples of successful initiatives

2.1 Crystallography

The role of crystallography bringing together biological and chemical 3 dimensional structure results have yielded treatments for HIV, and are actively engaged with current challenges like the zika virus. The biological and chemical crystal structure databases providing the durable archiving of precise and accurate data for new compound design is the scientific process involved firmly resting on across the disciplines data. These data are however united in their common type, derived from crystal structure analyses methods. These data also lend themselves to a reasonably homogenous description through the 'crystallographic information file' known as 'cif'. This firm ontology embedded in it, introduced after years of discussion in 1994, transformed the pace of archiving of crystallographic data in the databases such as the CSD and the PDB. It is possible that other across the disciplines data initiatives will present challenges of non-uniform descriptor types. Again in the health and diseases field the malaria challenge is a well-known one. Here the data on where mosquito nets are provided showed the most effective prevention of the spread of the disease although after infection the structure based drug design approach is being tried. That said the release of single sex, highly successful, male only versions of the tsetse fly may, like mosquito nets, prove the most effective. These data measuring success will be of various types and across a wider set of disciplines than biology or chemistry alone with the HIV example described above.

2.2 Bioinformatics

The fall in price of DNA and RNA sequencers in the last ten years has meant that life science is producing huge amounts of data. It is estimated that by 2020 life science data will be generated at a million times the current rate. Life science data in general, beyond genomics and genetics, exist in a wide range of formats, and is described in different ways. This makes it difficult to merge data sets and analyse the data. By themselves most research centres do not always have the facilities or the the expertise to store, manage, harmonize, share or analyze the data; or when systems are in place, these are fragmented and not always interoperable. Biological science increasingly involves large amounts of data and it is not easy to find the right resources: for example the appropriate software, the standards, the data repository, or the training material to get guidance and education.

To address this, ELIXIR has been established as the Europe-wide approach to bring a wide range of resources under the umbrella of the ESFRI (European Strategy Forum on Research Infrastructures). A Europe-wide approach is a first step towards a global solution which will be needed as Life Science becomes a Big Data science on a comparable scale to particle physics and astronomy. ELIXIR unites Europe's leading life science organisations in managing and safeguarding the increasing volume of data being generated by publicly funded research. It coordinates, integrates and sustains bioinformatics resources across its member states and enables users in academia and industry to access services that are vital for their research. ELIXIR includes 21 members and over 180 research organisations. It was founded in 2014, and is currently implementing its first five-year scientific programme. Each member provides data resources, tools and other services; for example FAIRsharing (formerly BioSharing) is a resource on standards provided by the ELIXIR-UK Node. This resource is further described in paragraphs 17-18 above .

2.3 Astronomy

Since ~1980 the astronomy community has shared a common data format known as the Flexible Image Transport System (FITS).[9] Begun as a way to exchange imaging data between optical and radio observatories, FITS has evolved to support many other types of data in astronomy and has even been adopted by the Vatican to store the digitized images of their cultural artefacts. FITS is officially endorsed by the International Astronomical Union, with changes and updates managed by Commission B2 on Data and Documentation. Astronomical data archives worldwide store data in the FITS format, and every major software package in the astronomy community has tools for reading and writing FITS data.

Building on the widespread use of FITS, the astronomy community began in 2001 to develop a distributed but federated system for data discovery, access, and interoperability. The Virtual Observatory (VO)[10] goes beyond the mostly syntactic standard of FITS to implement a semantic standard that allows for direct comparison of data from ground- and space-based observatories providing data across the electromagnetic spectrum. The VO's suite of data access protocols define a set of metadata elements that are sufficient to enable interoperability (e.g., image and spectral coordinate alignment and object cross-matching in non co-located databases). Data discovery is supported through a "resource registry" containing metadata about data collections and services. The VO standards are overseen by the International Virtual Observatory Alliance.[11] The VO architecture has been adopted in a number of other fields, from metrology and materials science to neuroscience.

2.4 Archeaology (Open Context)

Open Context (OC) is a free, open access online platform for researchers in archaeology and related disciplines, to electronically publish primary field data and documentation. It offers researchers various services to help them prepare and publish their data, such as web services and editorial review. The platform also serves as a portal for easy browsing and searching. The aim of OC is to make archaeological field data freely and easily accessible on the Web. Additionally, it wants to encourage data sharing and (re)use. It therefore strives to publish archaeological datasets as Linked Open Data, such that the data sets provided on the website can be easily referenced by unambiguous identifiers and they include links to other resources on the Web. OC is the result of a project, funded by the National Endowment for Humanities and the Institute of Museum and Library Services. Currently, OC is maintained and administered by the Alexandria Archive Institute40, a not-for-profit organisation41, based in Berkeley, California, while IT development is carried out in collaboration with the Berkeley School of Information. OC furnishes useful information regarding attitudes, practices and policies within the ecosystem of archaeology, as well as significant information regarding the technical approach adopted for the deposition of, accessibility to and preservation of the data it contains. OC relies on other repositories, like the California Digital Library (CDL) at the University of California, for the preservation of data and maintaining its quality. CDL, established in 1997, provides data archiving and curation services. Such services include persistent identifier services, data storage and guidance on data management planning. (Taken from the RECODE Final Report, p.24: http://recodeproject.eu/wp-content/uploads/2014/09/RECODE-D4.1-Institutional-barriers-FINAL.pdf)

2.5 Social Sciences (CESSDA and DDI)

The organisations that steward and provide access to digital data in the social sciences are relatively longstanding. The Inter-university Consortium for Political and Social Research https://www.icpsr.umich.edu/icpsrweb/ (which serves as the main archive for digital data in the social sciences for the USA) was founded in 1962, while the 'Social Science Research Council Data Bank', the forerunner of the UK Data Archive was established in 1967 http://data-archive.ac.uk/. These

[9] https://fits.gsfc.nasa.gov/
[10] Hanisch et al. 2015, http://dx.doi.org/10.1016/j.ascom.2015.03.007
[11] http://ivoa.net/

organisations, and their counterparts in other countries, provide essential services without which some areas of social science research would not be possible.

There is a recognised need for improved data discovery and cross searching for international comparisons and regional studies as well as acknowledgement that a lot of pertinent social science data is not as accessible and reusable as it might be, notwithstanding necessary restrictions where they exist. CESSDA ERIC https://www.cessda.eu/, which is the evolution of the Consortium of European Social Science Data Archives through the ESFRI processes into the ERIC structures, is an ongoing attempt to address this on a regional, European level. Coordination and collaboration of European social science data archives has as its objective better to 'researchers' access to important resources of relevance to the European social science research'. This requires ongoing work 'to develop and coordinate standards, protocols and professional best practices pertaining to the preservation and dissemination of data and associated digital objects'. The vision is to achieve this through by providing 'a full scale sustainable research infrastructure.' https://www.cessda.eu/

The data standard for much social science research is DDI, the Data Documentation Initiative and is curated by the DDI Alliance http://www.ddialliance.org/. DDI is 'describing the data produced by surveys and other observational methods in the social, behavioural, economic, and health sciences'. The full DDI standard takes a research lifecycle approach https://www.ddialliance.org/training/why-use-ddi and links to other standards http://www.ddialliance.org/standards/relationship-to-other-standards. The DDI Alliance has initiatives to develop further controlled vocabularies and RDF vocabularies to facilitate 'identifying programmatically the relevant datasets for a specific research purpose' http://www.ddialliance.org/Specification/RDF. The use of the Internet of Things / Sensor Networks, transactional data, social media data offers prospects that are both exciting and technically and ethically challenging for social science research.

2.6 Earth Science (OneGeology)

Traditional geology has some significant advantages over other disciplines in that the basic conceptual model, exemplified by the geologic map with its 'mapped units' and standard structures (folds, contacts, faults and fabric elements), has been essentially stable since Smith's map of Britain over 200 years ago, and the global process framework since the plate tectonics revolution of the 1960s. In addition, there has been exemplary institutional stability and homogeneity, in the form of 'geological surveys' which have existed in a similar form for decades in most national and many sub-national jurisdictions, and who publish scientific datasets that provide a basis for research projects in academia. There is also a community tradition of collaboration in the public sector, partly driven by the need to align scientific data across jurisdictional borders which rarely coincide with geological boundaries. Building on this, the OneGeology project was based on a consortium of more than 100 primarily national geological surveys. OneGeology developed standards for encoding and transfer of geologic map data to support a digital representation of global geology, at a broad scale at least (GeoSciML), with the data combined from live feeds from the original custodians. As well as the community scientific information model, OneGeology also relied on data and transfer standards coming out of the geospatial data community (OGC's WMS, WFS and Geography Markup Language), which was originally underwritten by investment from the defence sector (particularly in USA) as well as trans-national concerns about environmental information (in Europe, Canada, Australia).

The IUGS, through the Commission of the management and application of Geoscience Information (IUGS-CGI) was formed in 2004 to develop international data transfer standards for geoscience information. It now operates three standards working groups: 1) GeoSciML (in collaboration with the Open Geospatial Consortium (OGC)); 2) EarthResourceML and 3) Geoscience Terminology. GeoSciML is the standard for geological spatial data and sampling and has been ratified as an OGC standard. EarthResourceML is for the interchange of Earth resources data such as mineral deposits, mining activity and mining waste. The Geoscience Terminology group involves all the vocabularies that are required to

support data provided by the GeoSciML and EarthResourceML standards. The terms have definitions with source notes and are hierarchically organised to enable searching at different levels of granularity. Multilingual versions of the vocabularies are being planned.

In 2017 a proposal was released to develop an OGC GeoScience Domain Working Group. It aims to connect people interested in this topic to develop, improve and promote technologies for GeoScience data description and sharing. This working group is to be hosted by the OGC and co-chaired with CGI / IUGS. The GeoScience Domain Working Group will coordinate efforts with other OGC Earth science Domain Working Groups (agriculture, hydrology, etc.) under the umbrella of the Earth System Science DWG. A link with other 3D related working groups will also be developed (3DIM, Land and Infrastructure, Smart Cities, and more).

The Ocean Drilling Project (and successors), and the earth observations systems are both examples of science that depends on expensive-, and therefore shared-platforms, which more or less enforces the use of data standards and a model regime of data sharing.

2.7 Nanotechnology

Nanomaterials are complex, and researchers continue to develop new and innovative materials. Describing nanomaterials is a challenge for all user communities, but a description system is essential to ensure that everyone knows exactly which nanomaterial is being discussed, whether for research, regulatory, commercial, or other purposes. CODATA and VAMAS, an international pre-standardization organization concerned with materials test methods, have set up a joint working group to help develop a uniform description system for nanomaterials. This international working group includes representatives from virtually every scientific and technical discipline involved in the development and use of nanomaterials, including physics, chemistry, materials science, pharmacology, toxicology, medicine, ecology, environmental science, nutrition, food science, crystallography, engineering, and more. Fourteen international scientific unions actively participate.

One result of the working groups effort is the Uniform Description System for Materials on the Nanoscale (UDS). The UDS contains 19 tables of detailed descriptors and their definitions that are directly applicable for reporting nanomaterials research results, identifying nanomaterials in regulations and standards, developing formats for nanoinformatics resources, specifying nanomaterials in commercial transactions, and other uses. The UDS is now being considered for recognition as an international standard.

2.8 A Trans-disciplinary success

An example of transdisciplinary science that was made possible through standards and innovative informatics solutions. E.g. of eBird initiative, involving interdisciplinary data and citizen science. Specifically, eBird has been successfully used to identify bird migratory pathways globally, document species distributions, assess risks, and pinpoint critical ecosystems. eBird relies upon data generated by tens of thousands of citizen scientists as well as data derived from remote sensing, climate monitoring networks and other sources. Data challenges that eBird resolved included standardizing data and metadata formats, assuring the quality of data collected by tens of thousands of citizen scientist volunteers, integrating and analyzing multidisciplinary data collected across diverse scales of space and time, and developing new statistical and visualization approaches to analyze and visualize the enormous volumes of data. Sociocultural challenges addressed by eBird included engaging, training and rewarding a large “army” of citizen scientists, building a quality assurance program that engaged both human experts and automated statistical approaches, and working across many science domains and agencies.