Data Sharing and Interoperability

Francoise Genova

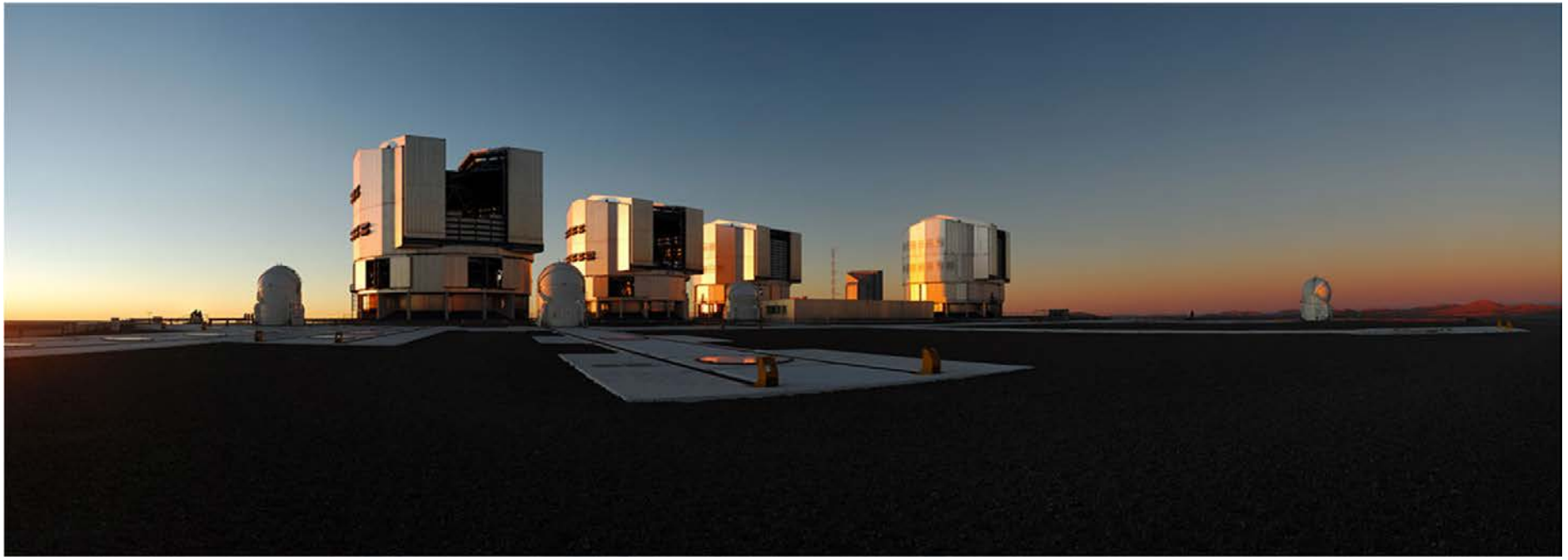
RDA TAB and RDA/Europe member

WDS Scientific Committee

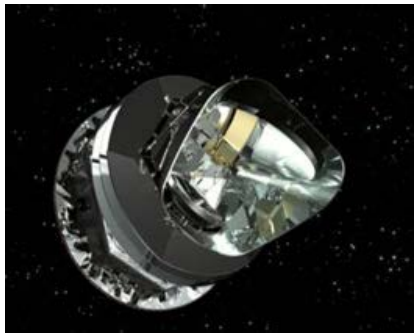
- *The general context in which DataCite works is a complex system with many elements*

- *A researcher's point of view on requirements*

How do astronomers do science?



And also …



We use research infrastructures

« Research infrastructures are facilities, resources and services that are used by the research communities to conduct research and foster innovation in their fields. Where relevant, they may be used beyond research, e.g. for education or public services . »

Definition from HORIZON 2020 Work programme 2014 – 2015

Data is, or should be, among the research infrastructures .

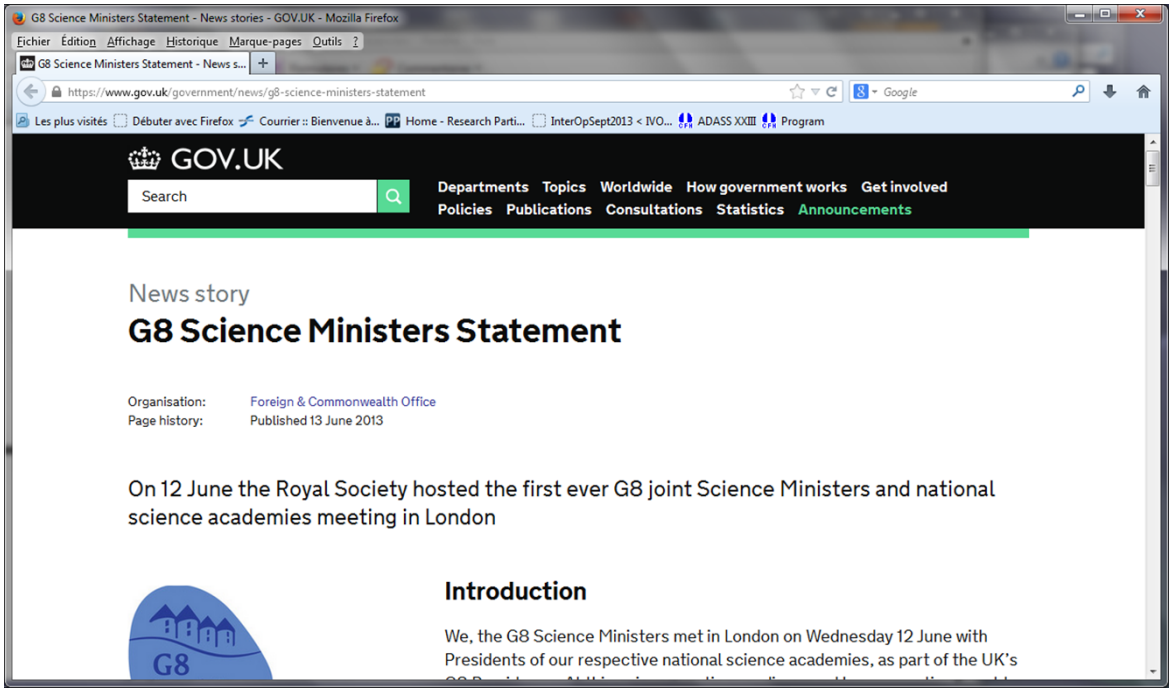
From the 2030 Vision of the « Riding the Wave » report to the European Commission (2010):

All stakeholders, from scientists to national authorities to the general public, are aware of the critical importance of conserving and sharing reliable data produced during the scientific process.

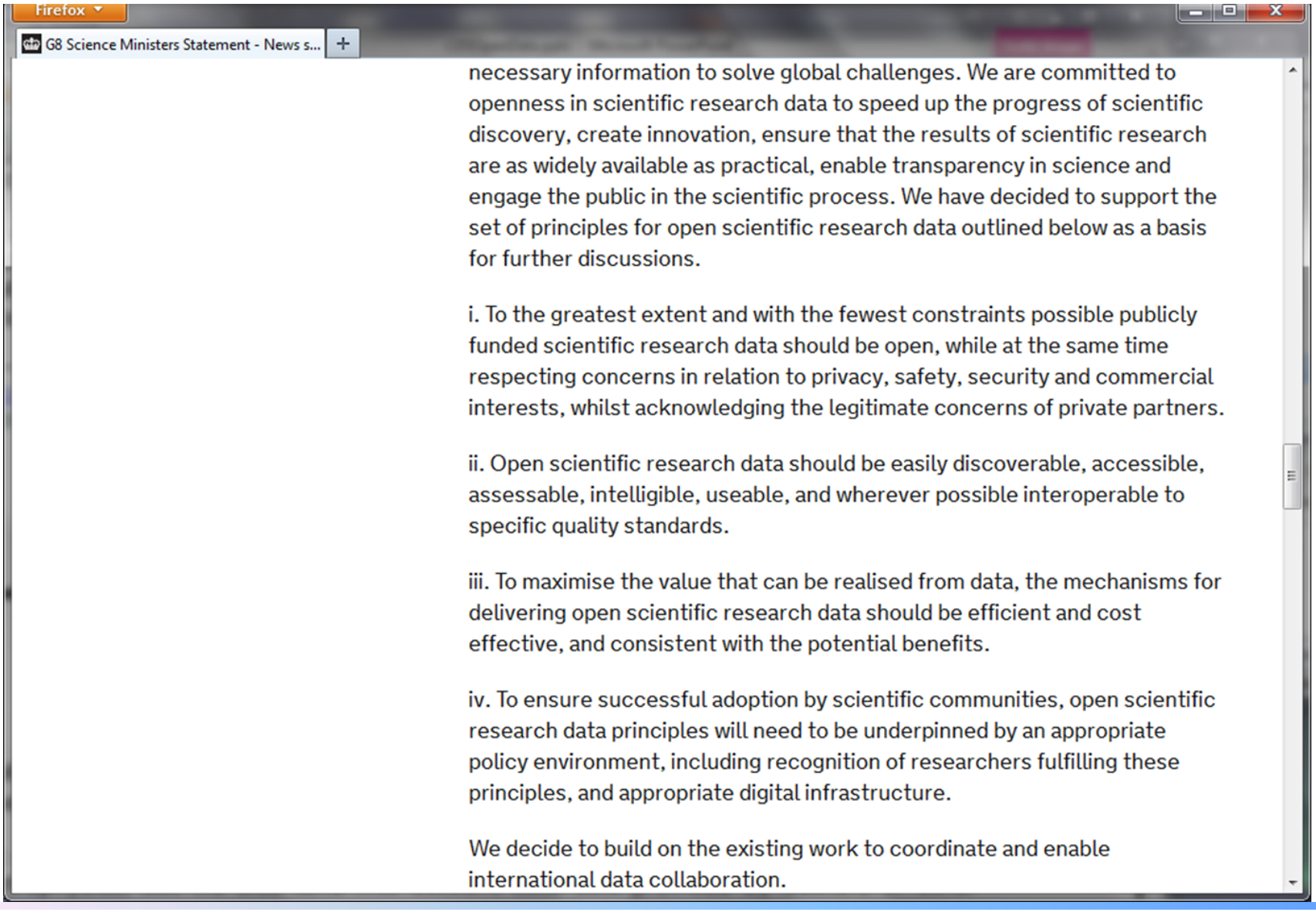
Impact if achieved

Data form an infrastructure, and are an asset for future science and the economy.

- This is true for all sciences, Social Sciences and Humanities as well as for sciences which use « physical » infrastructure
- Sharing of result/project data becomes a requirement on projects from many funding agencies
- « Open science » is currently a buzz word, even for the G8 Ministers of Science



« Open Science » in the G8 Ministers' statement



What we learnt in astronomy

- It is worth doing the job!
- In astronomy, the change of paradigm is done and astronomers use remote data in their everyday work – they probably could not live without it any more.
  - Key for main scientific subjects: multi-wavelength data to understand the physical phenomena at work in the objects, variable phenomena, etc…
  - Many more papers from data retrieved from observatory archives than from the original observations (HST, …)
  - Also huge usage of the value-added services, eg ~1 000 000 queries/day on the CDS services which are only one of the on-line data resource
- It is worth but it requires lots of work and takes a fraction of the community resources .

- It is not enough to set rules about data sharing, scientists have to be convinced to participate themselves and to give time and resources.
- How to build this willingness to extend data sharing to more/all disciplines?

Convince Data Producers/Users

- Data producers include
  - « Physical » Research Infrastructures
    - Although it is more and more often mandatory for Research Infrastructures to share their data, it is not always done and some have a hard work convincing their users who sometimes prefer to keep their data for themselves indefinitely or for a long time to avoid competition with others in the data usage.
    - Fairness on the provider and user sides:
      - allow a 'short' proprietary period e.g. when the data is obtained from a highly competitive process (one year in general in astronomy), also to continue to feed the system with 'the best' data
      - make sure that the data is not kept hidden forever or for too long
    - They are also a target in addition to research tems/researchers (cf. ARGO talk)
  - Individual scientists/research teams
    - Keywords: trust, usability of the data sharing process, usefulness of the data infrastructure, incentive and rewards, support to the data sharing process

The elements of trust

- The data system fulfils the user needs – a user-centered system, not a system driven by technology or by data preservation.
- Data must be usable: it must come with the metadata and in a format allowing reuse.

- Metadata, data formats, include elements which depend on the discipline.
  - Disciplines have to care about their data to set up the disciplinary components of data sharing (metadata et al.)

- Data providers must trust the data repository and more generally the data framework.
  - The data will be preserved on the long term > sustainability
  - Other people will be able to find and use the data
  - They are confident that data will be used, i.e. the data repository and the data framework more generally have the capacity to fulfill the research needs

  - Usefulness of certification
    - Certification criteria are a good topic for discussion to include the different possible points of view (cf the RDA-WDS WG)! Not only technical aspects, all the components of trust have to be identified.
  - E.g., do users want the original data or data corrected from errors? Preservation vs usage

Incentives and rewards

- Make it easy: support researchers in the process of sharing their data
  - Data repositories work at facilitating the deposit process
  - Support also needed to prepare metadata
  - Librarians in institutional repositories
  - *Disciplinary data repositories* (librarians, researchers, s/w engineers) gather the expertise and can play a major role
- Make it worth
  - When a 'good' data system is set up, a virtuous loop: when researchers use data from the system for their own research they are more eager to share their own one (although not always, in which case rules can be very useful)
  - Increase trust in one's own results by providing access to data allowing to check them
  - Increase one's research impact by allowing data reuse by others
  - Include data sharing among researchers' evaluation criteria

Evaluation…

- The fact that researchers' evaluation criteria have to evolve appears prominently every time « open data » is discussed, at all levels, e.g.
  - One of the first comments from the audience in any talk on data sharing
  - « To ensure successful adoption by scientific communities, open scientific research data principles will need to be underpinned by an appropriate policy environment, including recognition of researchers fulfilling these principles »

    G8 Science Ministers Statement, 2013
- Unfortunately nothing changed until now in the « bibliometry dictatorship » although most people understand the problems (take into account all types of research activities is only one of them)
- Recognize data sharing can also include measure data value

  « If we are to encourage broader use, and re-use, of scientific data we need more, better ways to measure its impact and quality »

  "Riding the Wave" report

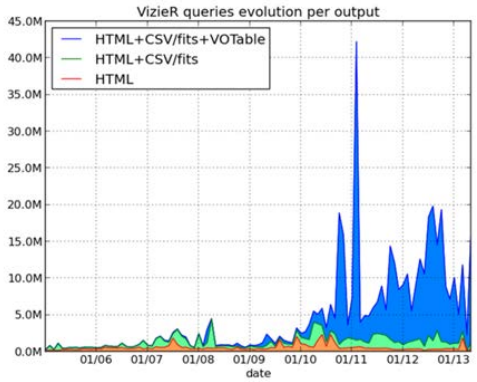
Data citation, data usage

- Of course here the capacity to cite data plays a major role
  - Ensure fairness by referring to the data used in a paper (like one expects researchers to cite the articles they use)
  - Enable data citation metrics to include in the evaluation criteria (but…)
  - Are they other criteria to measure data impact?
- This is also an important incentive for the Research Infrastructures to share their data
  - One aim is to increase their impact by allowing data reuse by others
  - It is necessary then to be able to identify publications using their data even if they are not from the original infrastructure user (author's name not known a priori)

To get the full benefit of data sharing

- *Discoverable, accessible, assessable, intelligible, useable and wherever possible interoperable data* (G8 Science Minister Statement)
- Interoperability among disciplinary resources

  VizieR usage through the VO protocols



- Increases the impact of data sharing, eg J/ApJ/594/1
  - Paper cited 1357 times since 2003, ~100 times in 2013 (ADS)
  - 1850 queries in VizieR in 2013, 93% through VO protocol from a VO tool

Interoperability

- Cross-discipline interoperability essential to tackle the grand challenges
- Generic building blocks can help many
- For each building block of the data infrastructure, there will likely be more than one solution, e.g. data citation
- Some disciplines already have well established systems and cannot be required to break something which works
- Interoperability is also the key here

The Research Data Alliance

- Breaking barriers to data sharing – also many scientific disciplines represented
- Bottom-up work to provide building blocks, best practices, etc.
- Working Groups such as
  - Data Citation WG *Making dynamic data citeable*
  - Data Type Registry WG
  - RDA/WDS Publishing Data: Bibliometrics
  - RDA/WDS Publishing Data: Services
  - RDA/WDS Publishing Data: Workflows
- Interest Groups such as
  - Education and training on handling of research data IG
  - Long Tail of Research Data IG
  - PID IG
  - RDA/WDS Publishing Data IG
  - Plenary 4 September 22-24 in Amsterdam, co-located ODIN ORCID and DataCite event

RDA Research Data Sharing without barriers
RESEARCH DATA ALLIANCE