Data as an infrastructure: CDS, astronomy and beyond

Françoise Genova

Sharing scientific data – Open science

- A « hot » topic, up to the political agenda
- G8 Science Ministers
  - Strong statements on Open Scientific Research Data, 13 June 2013

    https://www.gov.uk/government/news/g8-science-ministers-statement
  - Global Research Infrastructures include data sharing aspects, 9 October 2015, Berlin

    http://www.g8.utoronto.ca/science/2015-berlin.html#gris
  - Open Science – Entering a new era for science, 17 May 2016, Tsukuba

    http://www.g8.utoronto.ca/science/2016-tsukuba.html

The June 2013 G8 Set of principles

- i. To the greatest extent and with the fewest constraints possible **publicly funded scientific research data should be open**, while at the same time respecting concerns in relation to privacy, safety, security and commercial interests, whilst acknowledging the legitimate concerns of private partners.
- ii. Open scientific research data should be easily **discoverable, accessible, assessable, intelligible, useable, and wherever possible interoperable to specific quality standards**.
- iii. To maximise the value that can be realised from data, the mechanisms for delivering open scientific research data should be efficient and cost effective, and consistent with the potential benefits.
- iv. 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, and **appropriate digital infrastructure**.

Additional G8 Ministers' statements

- Following a report of the « Group of Senior Officials on Global Research Infrastructures » (GSO) (Oct. 2015)
  - Further progress on sharing and managing scientific data and information should be achieved, especially by continuing engagement with community based activities such as the Research Data Alliance RDA.
  - We encourage the GSO to continue their work on convergence and alignment of inter-operable data management that could accomplish an effective open-data science environment at the G7 level and beyond.
- Open Science statement – Entering into a new era for science (May 2016)
  - Establish a working group on open science with the aims of sharing open science policies, exploring supportive incentive structures, and identifying good practices for promoting increasing access to the results of publicly funded research, including scientific data and publications, coordinating as appropriate with the Organisation for Economic Co-operation and Development (OECD) and Research Data Alliance (RDA), and other relevant groups; and
  - Promote international coordination and collaboration to develop the appropriate technology, infrastructure, including digital networks, and human resources for the effective utilization of open science for the benefit of all.

Recent advances

- The FAIR Guiding Principles for scientific data management – Findable, Accessible, Interoperable, Reusable – already present in astronomy for a loooong time

  http://www.nature.com/articles/sdata201618

- Rapid emergence of the Research Data Alliance (RDA)

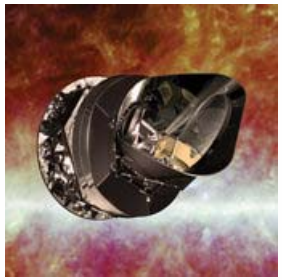
  https://www.rd-alliance.org

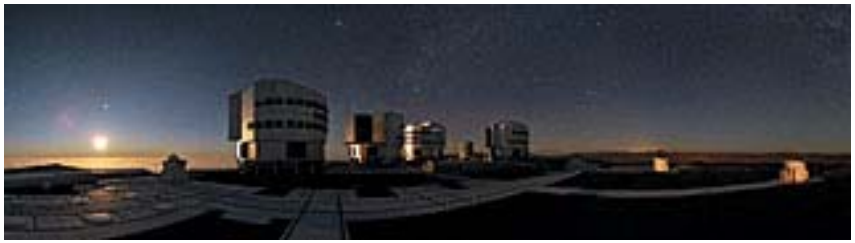
Not only a political subject!

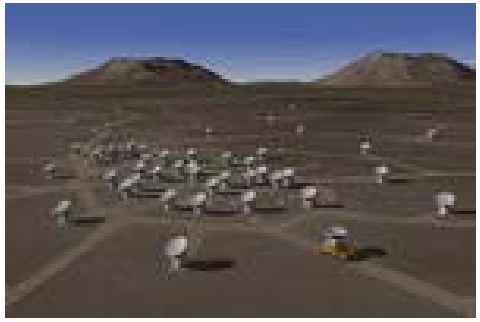
- A change in paradigm in the way science is done
- Astronomy as a case study

- In astronomy data is available AND USED!
  - More papers from data retrieved from HST archives than from original observations
  - 800.000 queries/day in average on the CDS services alone (only one element of landscape)

Astronomy Research Infrastructures

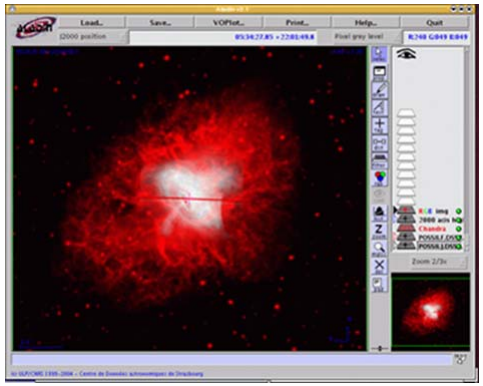


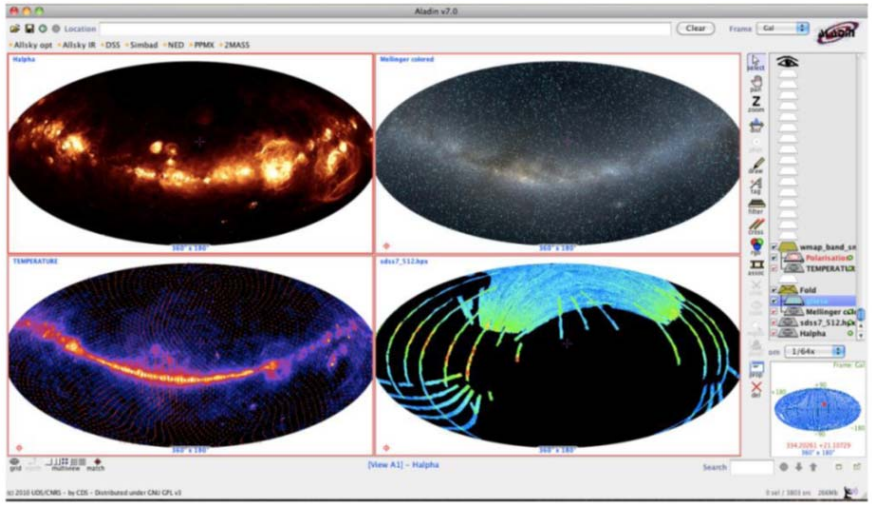




And data!

Why sharing data?





At the core of astronomy scientific needs !

- **Multi-wavelengths, multi-technique astronomy**
- **Time variability**
- **Comparison of theoretical models with observations**
- **Etc.**

Optimize the scientific return of the large infrastructures

How? Basic elements

- A common data format since the 70s (FITS)
- Strong tradition of international collaboration
- Open data in general (often after a proprietary period)
- Driven by community needs (on-line observation archives, on-line services)

Astronomical data

- Observations from ground- and space based telescopes (in general competitive calls for proposals)
- Sky surveys (homogeneous data set with up to billions of objects, measurements, images, spectra, time series)
- Modelling results
- Data from publications
- Value-added data bases, which gather homogenized information in particular from publications

  e.g. SIMBAD, names of objects and papers where the object is cited:

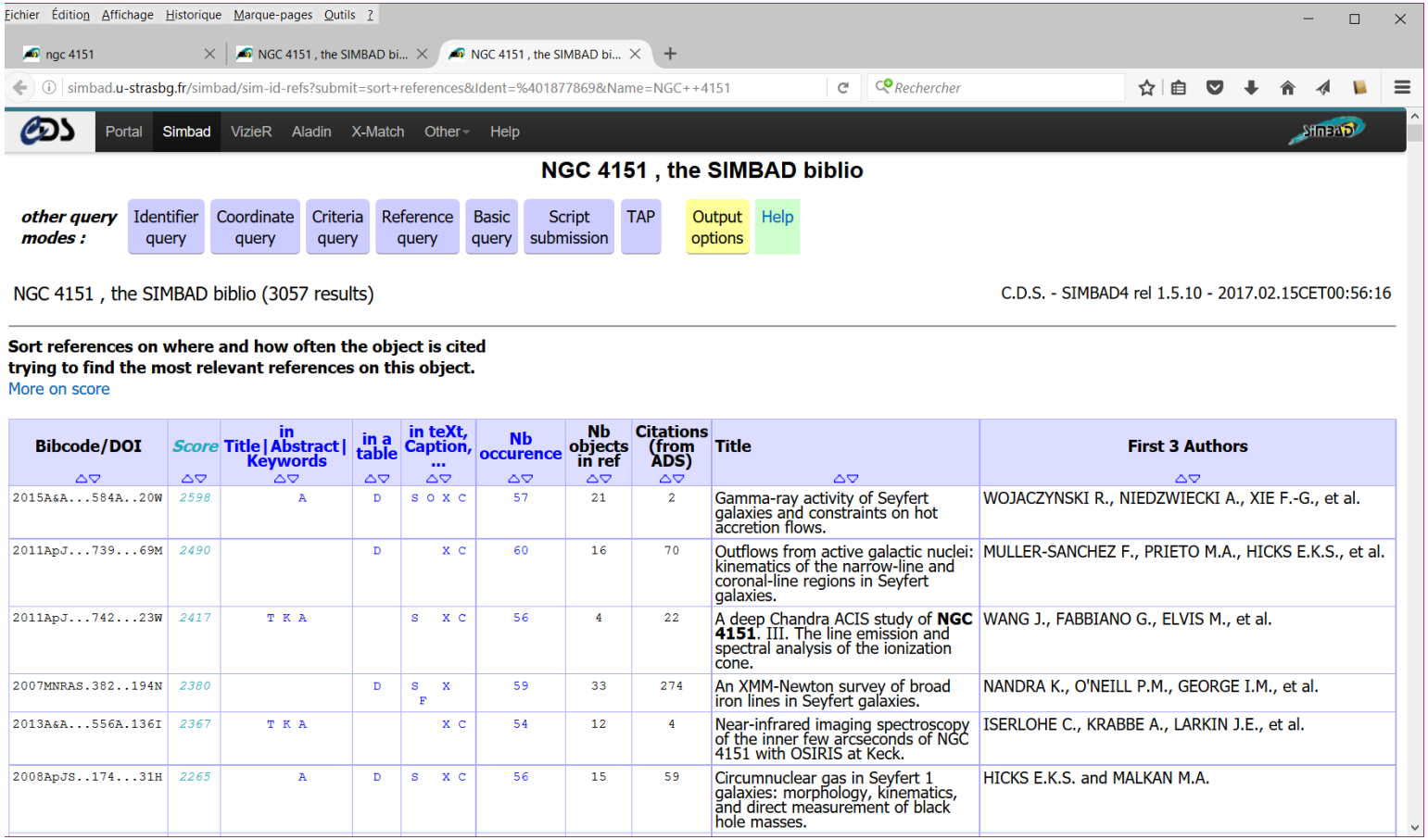
  9 200 000 objects, 24 700 000 object names, 330 000 references, begun ~1970

  also NED, VizieR, ADS for bibliographic data

Added-value databases : the example of SIMBAD

- Information from catalogues and publications
- Lots of work behind the scene (astronomers/computer engineers/specialized librarians)
- All the names of a given object
  - Used by archives (together with NED and VizieR) to transform names into coordinates – their access key
- The references in which the object was cited
  - Also information on relevance : where the name is cited in the paper

SIMBAD added value: references citing the object ⊕ sort references by relevance



Fichier Édition Affichage Historique Marque-pages Outils ?

ngc 4151 | NGC 4151 , the SIMBAD bi... | NGC 4151 , the SIMBAD bi...

simbad.u-strasbg.fr/simbad/sim-id-refs?submit=sort+references&Ident=%401877869&Name=NGC++4151

Rechercher

Portal | Simbad | VizieR | Aladin | X-Match | Other | Help

NGC 4151 , the SIMBAD biblio

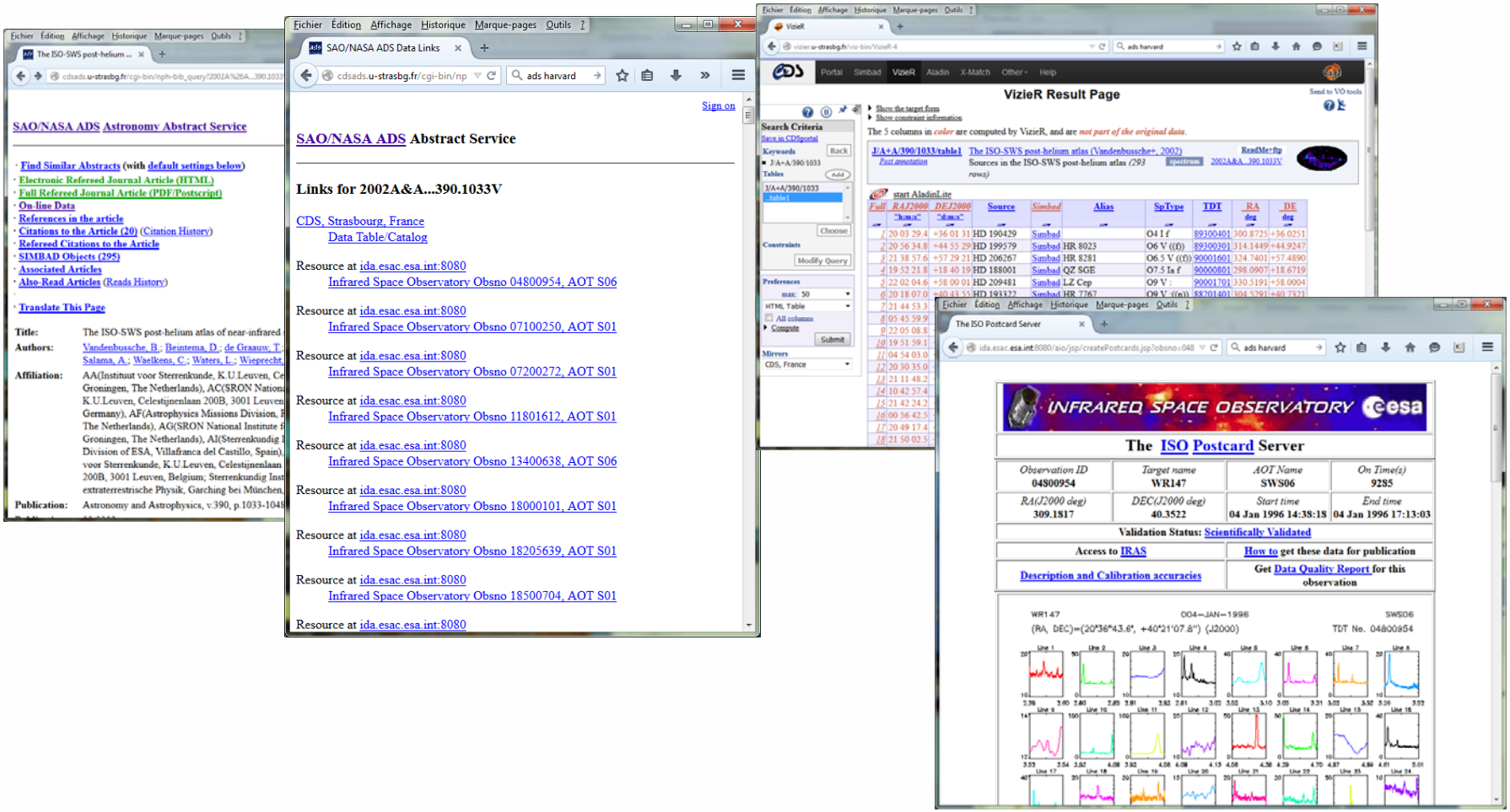
other query modes : Identifier query | Coordinate query | Criteria query | Reference query | Basic query | Script submission | TAP | Output options | Help

NGC 4151 , the SIMBAD biblio (3057 results) C.D.S. - SIMBAD4 rel 1.5.10 - 2017.02.15CET00:56:16

Sort references on where and how often the object is cited trying to find the most relevant references on this object.
More on score

Bibcode/DOI	Score	in Title | Abstract | Keywords	in a table	in teXt, Caption, ...	Nb occurence	Nb objects in ref	Citations (from ADS)	Title	First 3 Authors
2015A&A...584A..20W	2598	A	D	S O X C	57	21	2	Gamma-ray activity of Seyfert galaxies and constraints on hot accretion flows.	WOJACZYNSKI R., NIEDZWIECKI A., XIE F.-G., et al.
2011ApJ...739...69M	2490		D	X C	60	16	70	Outflows from active galactic nuclei: kinematics of the narrow-line and coronal-line regions in Seyfert galaxies.	MULLER-SANCHEZ F., PRIETO M.A., HICKS E.K.S., et al.
2011ApJ...742...23W	2417	T K A		S X C	56	4	22	A deep Chandra ACIS study of **NGC 4151**. III. The line emission and spectral analysis of the ionization cone.	WANG J., FABBIANO G., ELVIS M., et al.
2007MNRAS.382..194N	2380		D	S X F	59	33	274	An XMM-Newton survey of broad iron lines in Seyfert galaxies.	NANDRA K., O'NEILL P.M., GEORGE I.M., et al.
2013A&A...556A.136I	2367	T K A		X C	54	12	4	Near-infrared imaging spectroscopy of the inner few arcseconds of NGC 4151 with OSIRIS at Keck.	ISERLOHE C., KRABBE A., LARKIN J.E., et al.
2008ApJS..174...31H	2265	A	D	S X C	56	15	59	Circumnuclear gas in Seyfert 1 galaxies: morphology, kinematics, and direct measurement of black hole masses.	HICKS E.K.S. and MALKAN M.A.

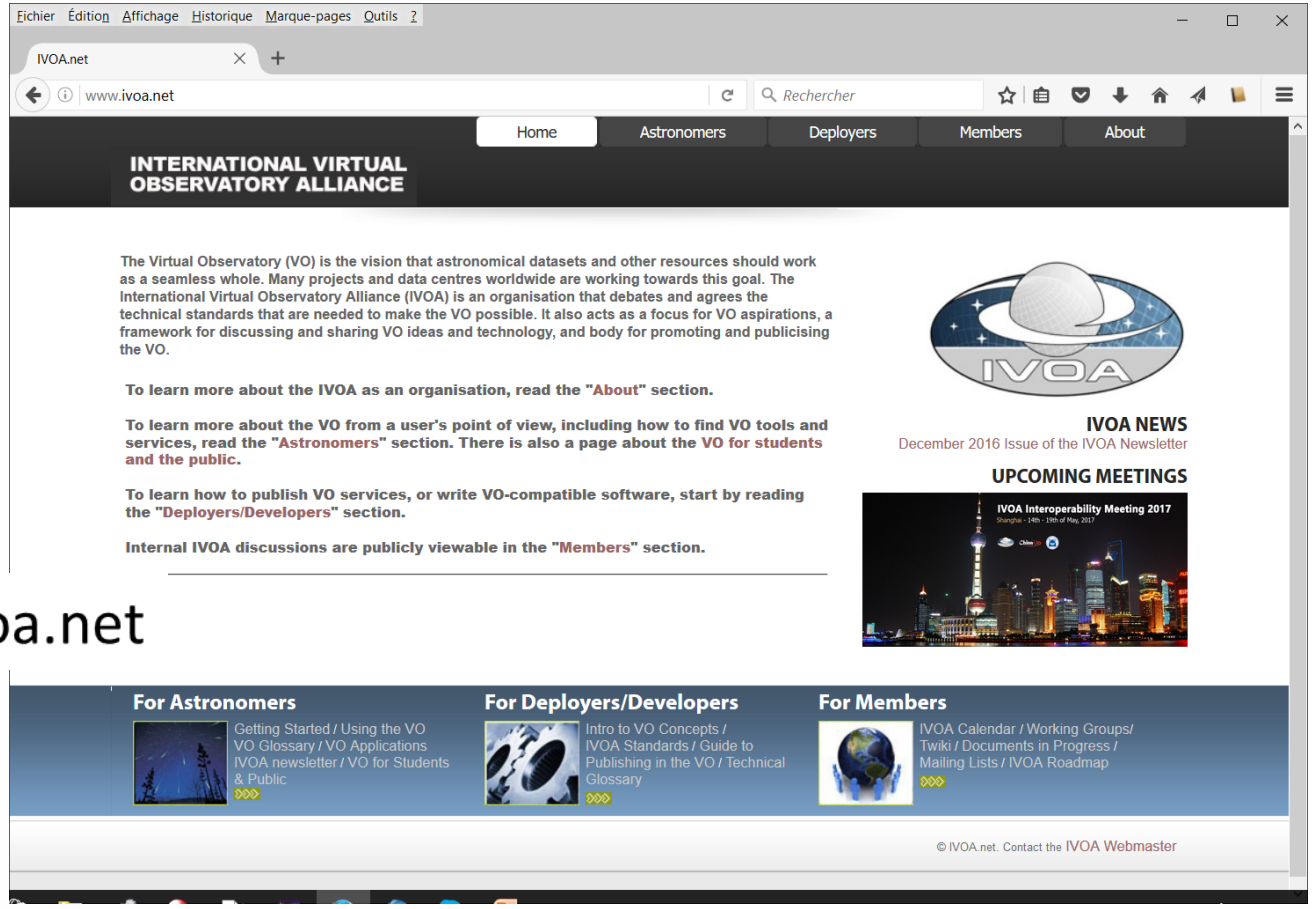
Early networking of on-line resources, still in use



Networking and interoperability

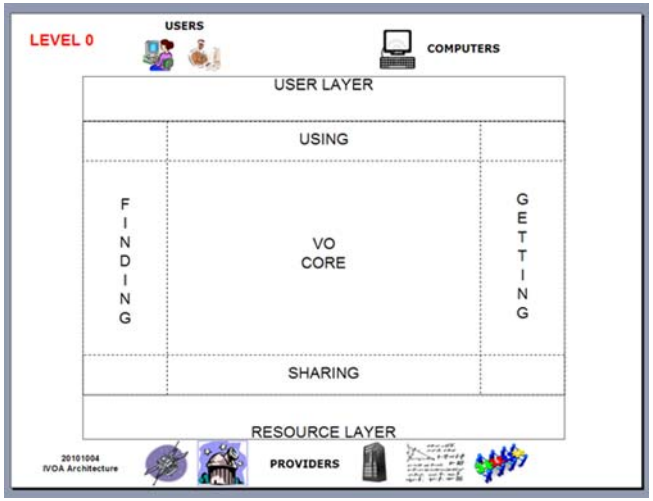
- **Networking of on-line resources from 1993-4 (added-value services, journals, archives)**
- **Seamless access to on-line data (~2000)**
  The astronomical Virtual Observatory
- **The VO framework : standards and data access tools – discover, access, use data**
- **Standards defined by the International Virtual Observatory Alliance (IVOA)**
  - **An alliance of national VO initiatives**
  - **Procedure inspired from W3C**
  - **When possible generic elements**
    (Registry: OAI-PMH, Vocabularies: SKOS/RDF)

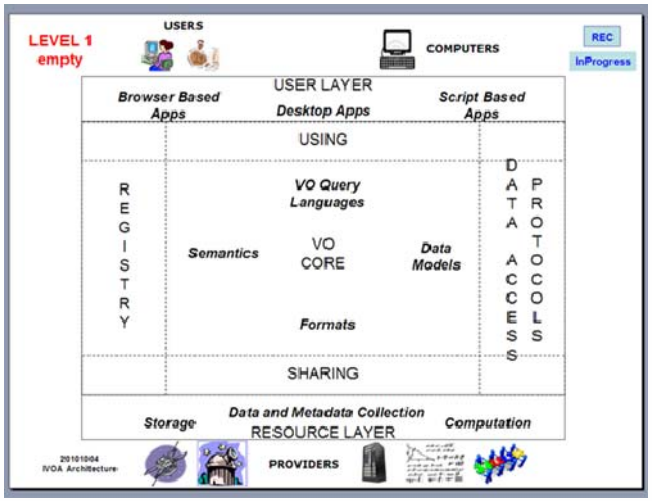
The IVOA

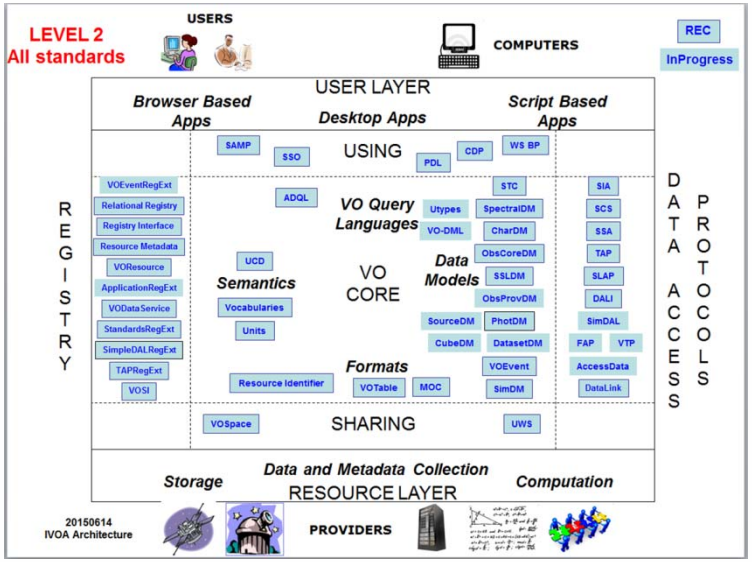


http://ivoa.net

The IVOA standard framework



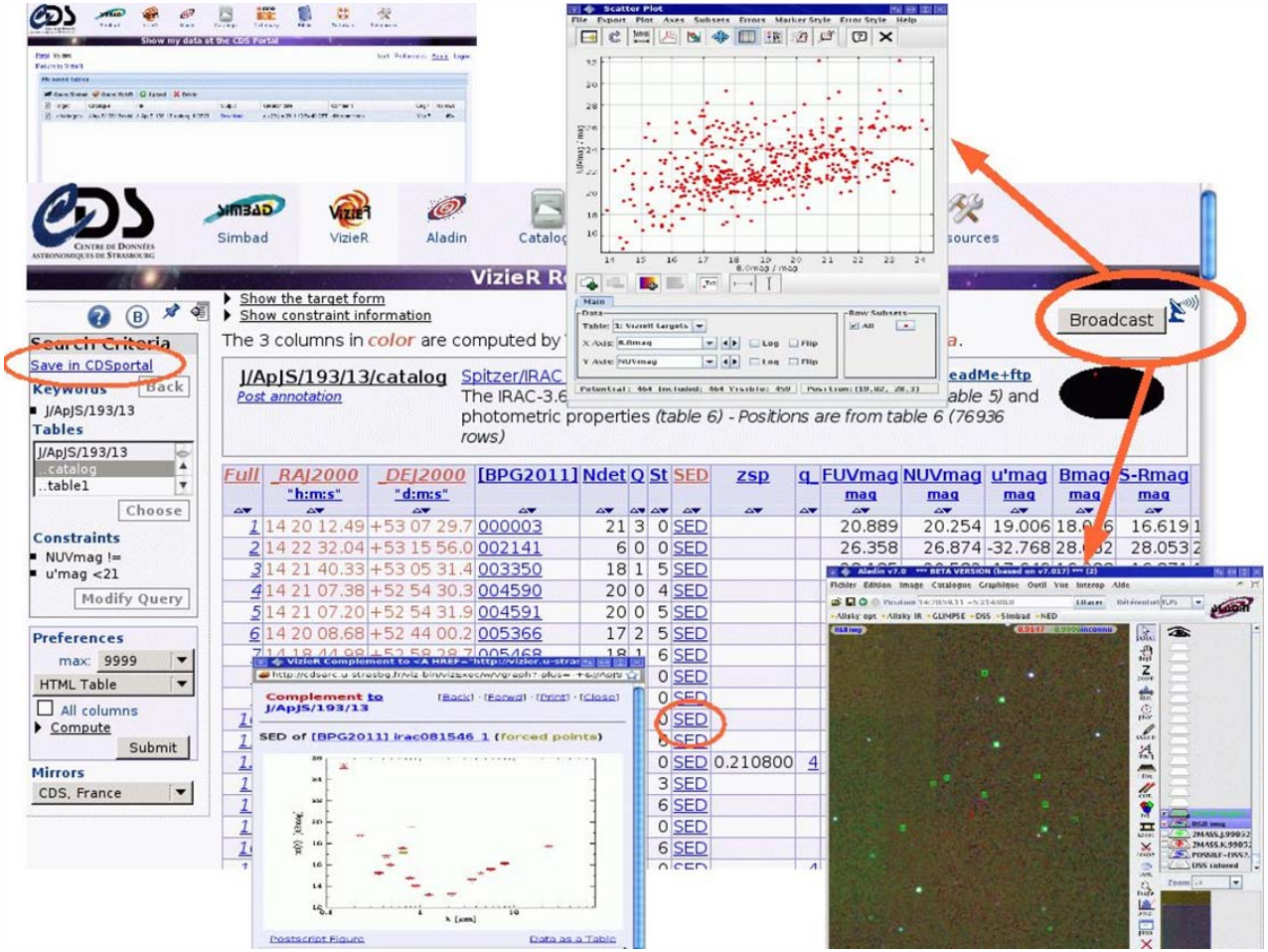




An inclusive and open framework

- **No central point, a multi-polar world, a global endeavour**
- **“Open” and inclusive model**
  - **A thin interoperability layer on top of the data holdings**
  - **Anyone can register a data service or build a tool (more than 100 “authorities” with a registered service)**
- **The VO is invisible but used - astronomers use the services and the tools!**
- **Not only the interoperability layer now: data providers also imbed VO building blocks in their archives and services**

Interoperable tools and data services



Keys for success

- **Key for success (science users): seamless access to data AND interoperable tools relevant to science needs**
- **Keys for success (data providers):**
  - **More visibility for their data**
  - **No need to reinvent the wheel, people already worked and propose solutions (for data sharing but also elements of archives and services systems)**

Current status

- The VO framework is operational and used
- Three pillars
  - Support to data providers
  - Support to science users
  - Technological work to update standards and tools

Genova et al. 2015

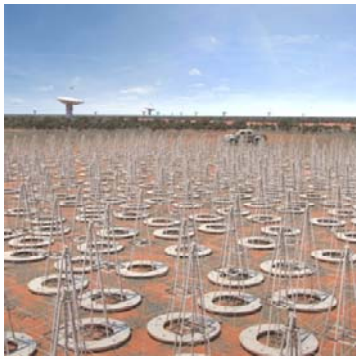
- Priorities linked to the needs of the future large projects
  - Multi-Dimensional data – first milestone done, May 2017
  - Time domain

Current step in Europe : large projects

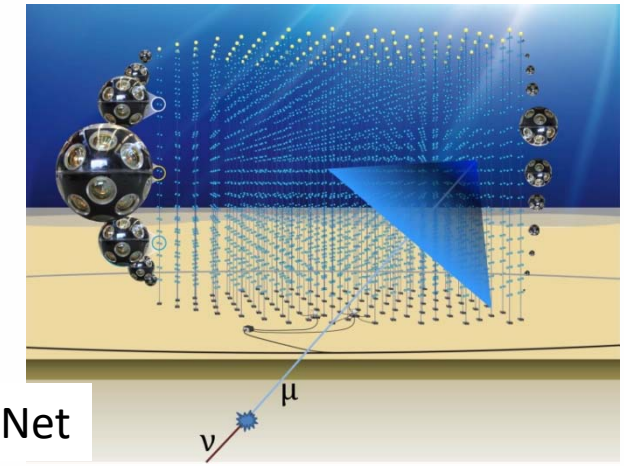
- ASTERICS WP4: Data Access, Discovery and interoperability (4.5 M€ on 4 years)
- “Make the ESFRI and pathfinder project data available for discovery and usage by the whole astronomical community, interoperable in a homogeneous international framework, and accessible with a set of common tools.”
- Fully aligned with the current IVOA priorities

Who is involved in ASTERICS WP4

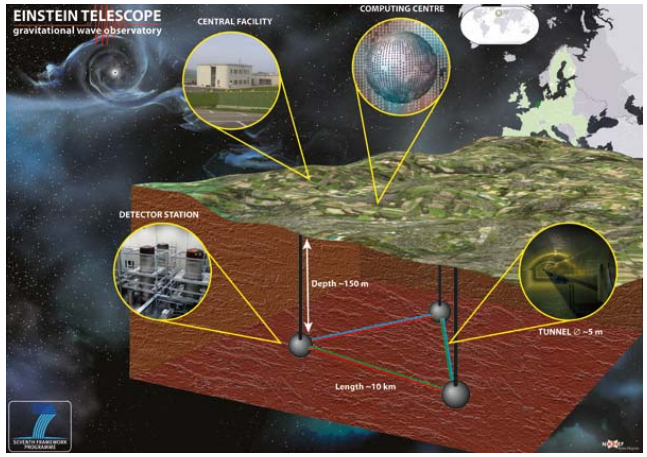
CTA



SKA



KM3Net



Asterics
Astronomy ESFRI & Research Infrastructure Cluster

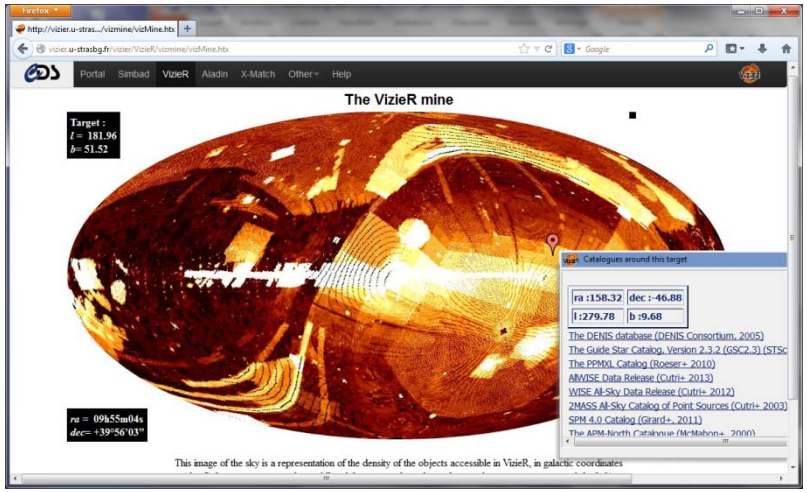
Who is involved in ASTERICS WP4

- Euro-VO partners, i.e. VO initiatives from France, Germany, Italy, Spain, UK
- Representatives of ESFRI and pathfinders
- Astronomy & Astroparticle physics, including new messengers
- ESO is associated to the project
- ESA (ESAC) is working in close collaboration with Euro-VO

Big and smaller data = astronomical data

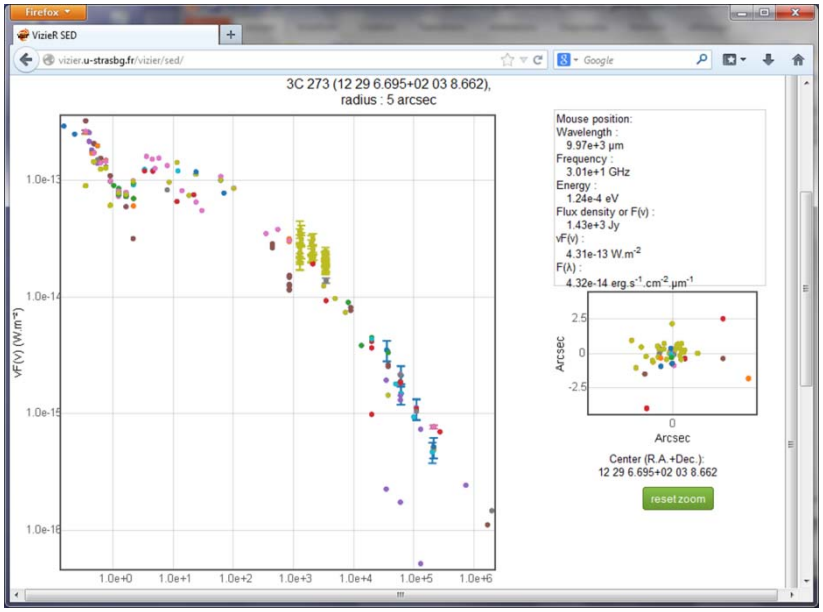
- **Observatory archives + disciplinary data centres**
- **Also data from publications – research results**
  - **Agreement between CDS and the journals (started in 1993 with *Astronomy & Astrophysics*)**
    - **tabular data from publications (also images, spectra, time series)**
    - **together with catalogues from sky surveys, space missions (up to 2 billion rows)**
    - **15 000 "catalogues", i.e. data sets**
  - **Homogeneous metadata describing the very heterogenous content**
  - **Fully discoverable, usable and used**
- **Not so much "big" and "little" data, but rather Useful, Validated and Documented data, huge diversity**

« Long Tail » data in VizieR



Data validated by a publication
Fully discoverable and usable
Together with the very large surveys

"Photometry viewer":
Spectral points extracted from the collection



Cross-disciplinary aspects

- Elements of the VO framework are customized and reused by « nearby » disciplines – planetary studies, the Virtual Atomic and Molecular Data Centre, also Materials sciences
- The generic elements (registry of resources, vocabulary concepts) allow astronomy data infrastructure to interface with the generic data framework

Data as a research infrastructure

- In Europe: Research Infrastructure Roadmaps
  - European level (ESFRI) and national Roadmaps
- Data/Computing/Network questions in the questionnaires of the ESFRI Roadmap and (e.g.) French National Research Infrastructure Roadmap
- Some of the Infrastructures in the Roadmaps are « virtual », dealing with data, others have a strong data component

Examples - Humanities

- ESFRI : DARIAH, CLARIN
- France: Huma-Num

  http://www.huma-num.fr/

  - Added-value services
  - Disciplinary Consortia to organize communities around data sharing topics

Example in France : Earth Sciences

- Disciplinary « Poles » with participation of all the organisations involved, incl. CNES
- « Inter-Pôle » technical discussions
- Overarching structure being built
- Strong links to European projectsin the different domains

Disciplinary Interoperability Frameworks

- Session at SciDataCon 2016
- Humanities/linguistics, astronomy, earth sciences, material sciences/crystallography
- Commonnalities
  - Must be science driven
  - Defining the discipline-specific part of the interoperability standards is mandatory but difficult
  - Share data AND applications
  - Incentives to data sharing is a key question
  - Social aspects more challenging that technical ones
- Governance is more diverse, linked to the discipline organisation and history
- Many sharable aspects – use the RDA for that!

The Research Data Alliance

- Founded in March 2013 by Australia, EC, and NSF and NIST
- ~5500 members from more ~ 120 countries
- Bottom-up work to tackle all the aspects of scientific data sharing, technological as well as « sociological »
- Have a look at

  rd-alliance.org and join!

RDA
RESEARCH DATA ALLIANCE

Vision

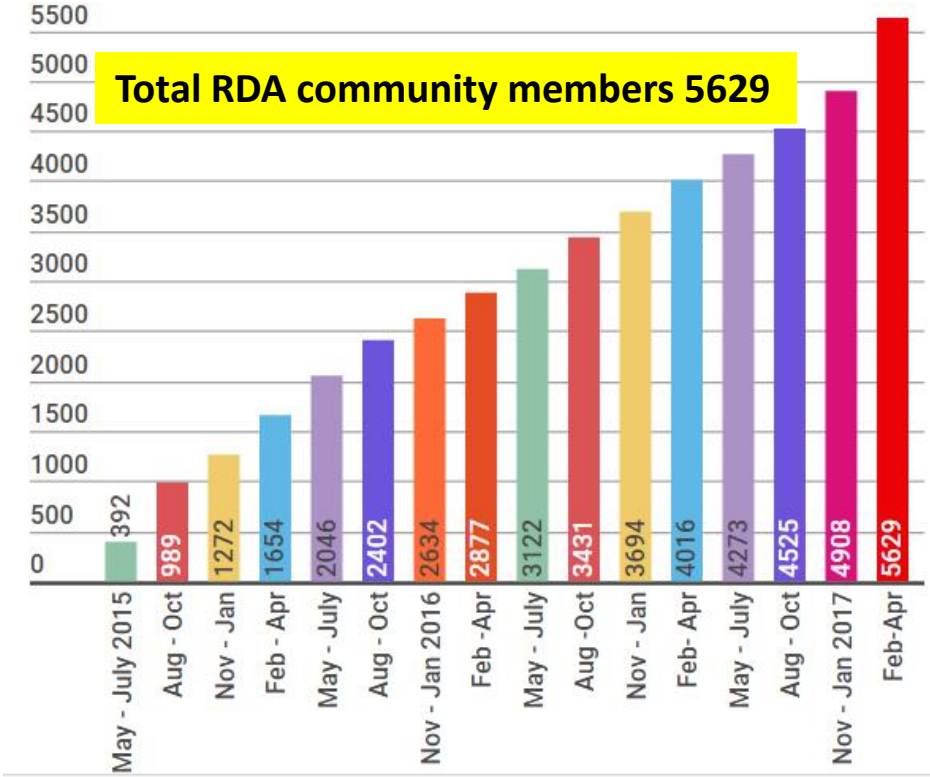
Researchers and innovators openly share data across technologies, disciplines, and countries to address the grand challenges of society.

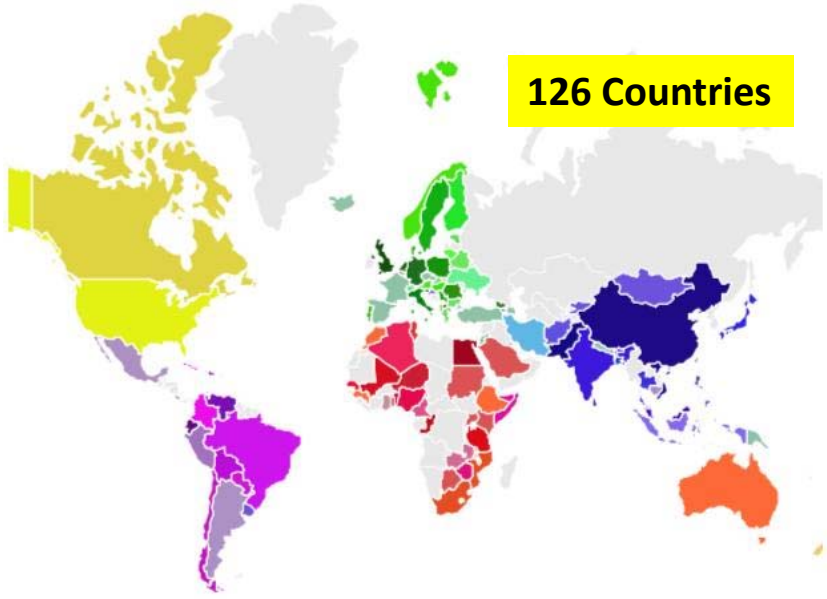
Mission

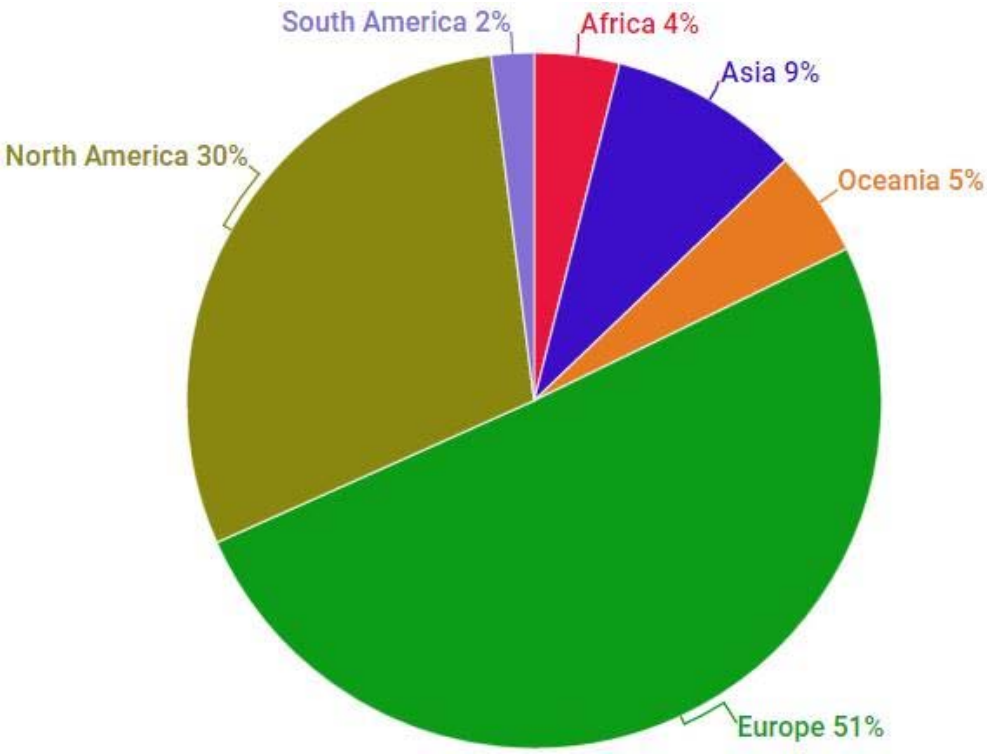
RDA builds the social and technical bridges that **enable open sharing** of data.

CC BY-SA 4.0

RDA worldwide growth







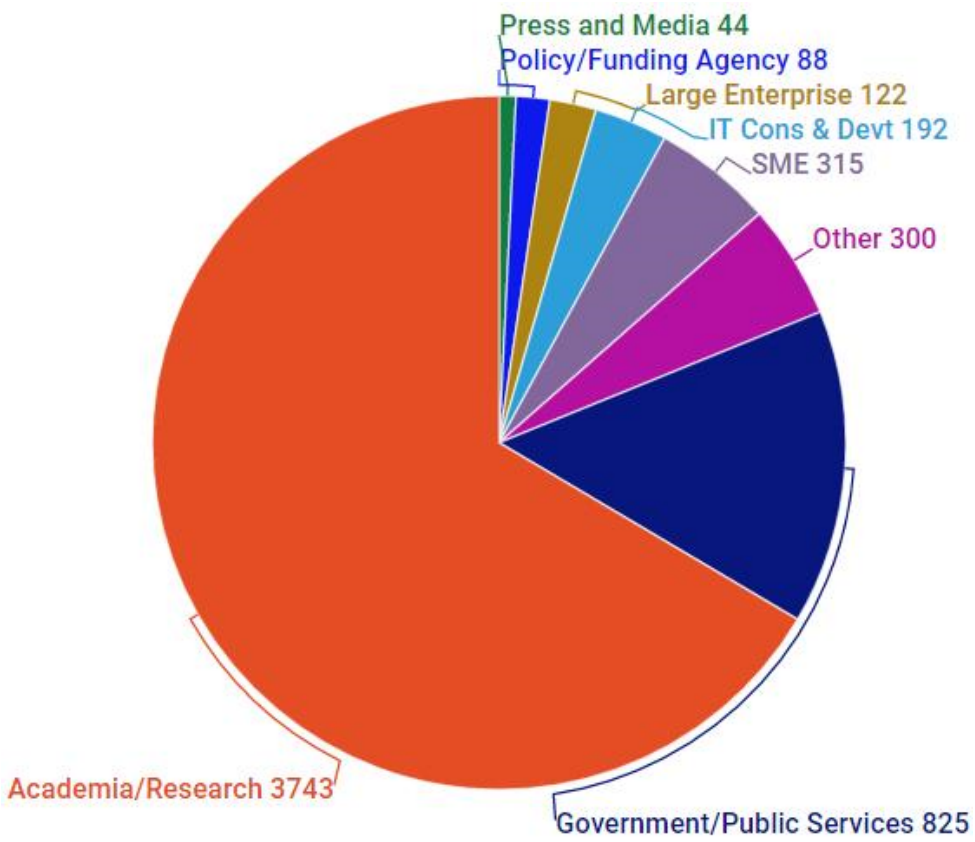
RDA
RESEARCH DATA ALLIANCE

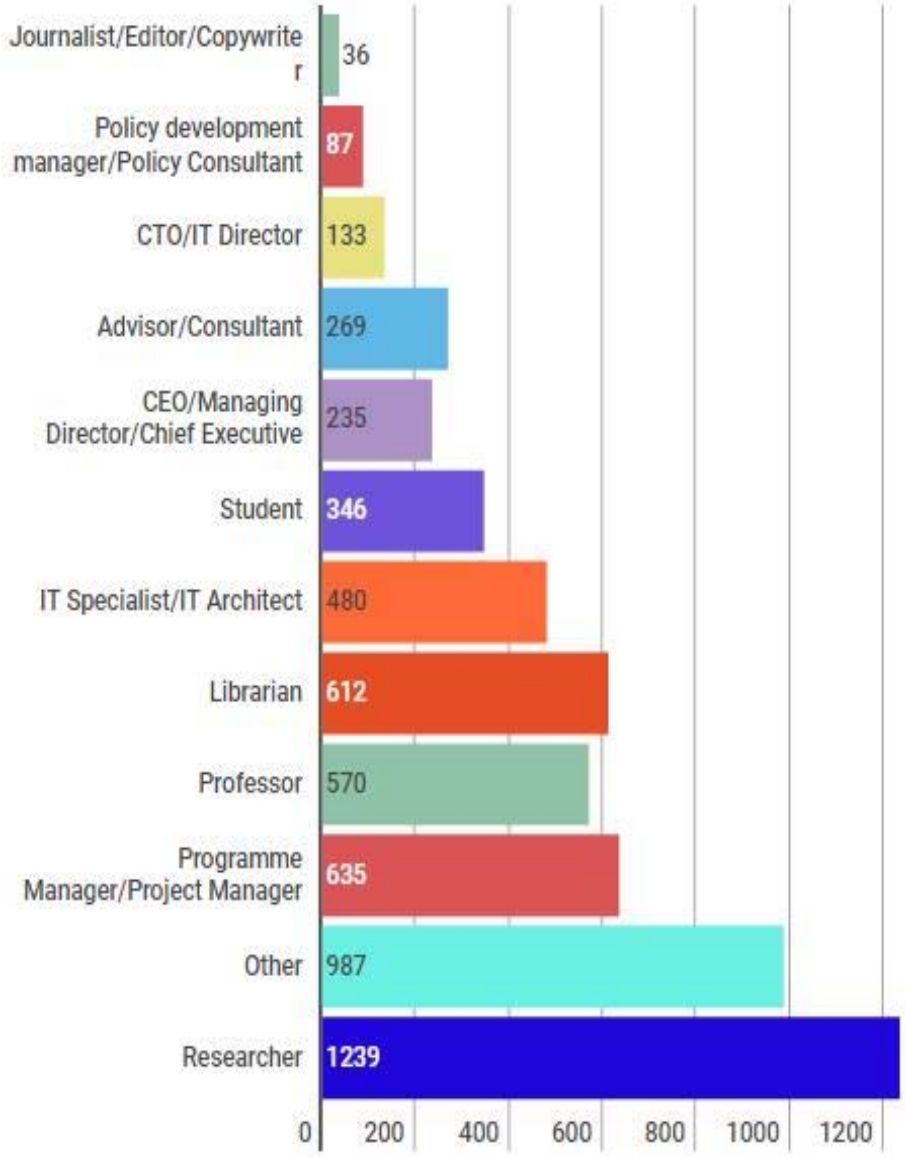
rd-alliance.org/about-rda

WWW.RD-ALLIANCE.ORG
@RESDATALL

CC BY-SA 4.0

Who is RDA?





CC BY-SA 4.0

Many topics, many participant profiles

- 82 Working Groups and Interest Groups on building blocks
- Wide diversity of topics
  - Domain science
  - Community needs
  - Data Referencing and sharing
  - Data stewardship and services
  - Base infrastructure
- Data providers, librarians, researchers, project/programme managers, publishers…
- International, neutral forum for discussion
- Recommendations and outputs

Among the opics of particular interest

- Repository audit and certification (DSA+WDS)
- Dynamic data citation
- RDA/WDS Publishing Data…
  - Bibliometrics
  - Workflows
  - Services (Scholix)
- 23 things: Libraries for Research Data
- Disciplinary Interoperability frameworks (including IVOA!)

Conclusion

- Exciting times for scientific data sharing
- Astronomy has been at the forefront
- Data is one of the discipline research infrastructures
- Global interoperability operational
- Many other disciplines are moving on
- The context is evolving very fast
- Important to participate in initiatives such as RDA to ensure our requirements are taken into account

RDA Interest (IG) & Working Groups (WG) by Focus (1)

Total 82 groups:
29 Working Groups & 53 Interest Groups

Domain Science - focused

- **Agrisemantics WG**
- **BioSharing Registry WG**
- **Fisheries Data Interoperability WG**
- **On-Farm Data Sharing (OFDS) WG**
- **Rice Data Interoperability WG**
- **Wheat Data Interoperability WG**
- Agricultural Data IG (IGAD)
- Biodiversity Data Integration IG
- Chemistry Research Data IG
- Digital Practices in History and Ethnography IG
- Geospatial IG
- Global Water Information IG
- Health Data IG
- Linguistics Data Interest Group
- Mapping the Landscape IG
- Marine Data Harmonization IG
- Quality of Urban Life IG
- RDA/CODATA Materials Data, Infrastructure & Interoperability IG
- Research data needs of the Photon and Neutron Science community IG
- Small Unmanned Aircraft Systems' Data IG
- Structural Biology IG
- Weather, Climate and air quality IG

Community Needs - focused

- **Certification and Accreditation for Data Science Training and Education WG**
- **RDA/CODATA Summer Schools in Data Science and Cloud Computing in the Developing World WG**
- **Teaching TDM on Education and Skill Development WG**
- Archives & Records Professionals for Research Data IG
- Data for Development IG
- Development of Cloud Computing Capacity and Education in Developing World Research IG
- Early Career and Engagement IG
- Education and Training on handling of research data IG
- Ethics and Social Aspects of Data IG
- International Indigenous Data Sovereignty IG

CC BY-SA 4.0

RDA Interest (IG) & Working Groups (WG) by Focus (2)

Total 82 groups:
29 Working Groups & 53 Interest Groups

Reference and Sharing - focused

- **Data Citation WG**
- **Data Description Registry Interoperability WG**
- **Data Security and Trust WG**
- **Empirical Humanities Metadata WG**
- **International Materials Resource Registries WG**
- **Provenance Patterns WG**
- **QoS-DataLC Definitions WG**
- **RDA / WDS Publishing Data Bibliometrics WG**
- **Repository Core Description WG**
- **Research Data Collections WG**
- **Research Data Repository Interoperability WG**
- Data Discovery Paradigms IG
- National Data Services IG
- RDA/CODATA Legal Interoperability IG
- Reproducibility IG

Partnership Groups

- **RDA / TDWG Metadata Standards for attribution of physical and digital collections stewardship WG**
- **RDA/WDS Scholarly Link Exchange Working Group**
- ELIXIR Bridging Force IG
- RDA/NISO Privacy Implications of Research Data Sets IG
- RDA/WDS Publishing Data IG

CC BY-SA 4.0

RDA Interest (IG) & Working Groups (WG) by Focus (3)

Total 82 groups:
29 Working Groups & 53 Interest Groups

Data Stewardship and Services – focused

- **Brokering Framework WG**
- **WDS/RDA Assessment of Data Fitness for Use WG**
- **RDA / WDS Publishing Data Workflows WG**
- Active Data Management Plans IG
- Data in Context IG
- Data Rescue IG
- Data Versioning IG
- Domain Repositories IG
- Libraries for Research Data IG
- Long tail of research data IG
- Preservation e-Infrastructure IG
- Preservation Tools, Techniques, and Policies IG
- RDA/WDS Certification of Digital Repositories IG
- RDA/WDS Publishing Data Cost Recovery for Data Centres IG
- Repository Platforms for Research Data IG
- Research Data Provenance IG
- Virtual Research Environments IG

Base Infrastructure – focused

- **Array Database Assessment WG**
- **Data Type Registries WG**
- **Metadata Standards Catalog WG**
- **PID Kernel Information WG**
- Data Fabric IG
- Data Foundations and Terminology IG
- Big Data IG
- Brokering IG
- Federated Identity Management IG
- Metadata IG
- PID IG
- Vocabulary Services IG

CC BY-SA 4.0