

# Lecture 4 – FAIR principles

## *Advanced Data Management*

Data Science and Scientific Computing / UniTS – DMG  
Scientific and Data-Intensive Computing / UniTS – DMG

Four guiding principles for scientific data management and stewardship

**F** indable

**A** ccessible

**I** nteroperable

**R** eproducible

Applicable to “data” but also to all what led to that data

- Algorithms
- Tools
- Workflows

# Added value repositories...



## Well-quoted

### Identifiers (32) :

An access of full data is available using the icon Vizier near the identifier of the catalogue

M 31	IRC +40013	NAME And Nebula	Z 0040.0+4100
2C 56	K79 1C	NAME Andromeda Nebula	[DGW65] 4
DA 21	LEDA 2557	NGC 224	[M98c] 004000.1+405943
2FGL J0042.5+4114	ZMASX J00424433+4116074	RAFGL 104	[VV2000c] J004244.3+411610
3FGL J0042.5+4117	ZMAXI J0043+412	UGC 454	[VV2003c] J004244.3+411610
GIN 801	MCG+07-02-016	UZC J004244.3+411608	[VV2006] J004244.3+411610
IRAS F00400+4059	NAME Andromeda	XSS J00425+4102	[VV2010] J004244.3+411610
IRAS 00400+4059	NAME Andromeda Galaxy	Z 535-17	[VV98c] J004245.1+411622

### Plots and Images

plot

radius  arcmin



CDS portal



CDS Simplay  
*(requires flash)*



Aladin applet

### References (10061 between 1850 and 2018) (Total 10061)

Simbad bibliographic survey began in 1850 for stars (at least bright stars) and in 1983 for all other objects (outside the solar system).

Follow new references on this object

Reference summaries :

from:  to:

or select by : (not exhaustive, [explanation here](#))

# General purpose repositories...



- Access
- No

All versions

Access Right

- Open (12)
- Closed (4)

File Type

- Pdf (9)
- Zip (2)
- Jpg (1)

Keywords

- Animalia (4)
- Biodiversity (4)
- Galaxies (4)
- Taxonomy (4)
- Astronomy (3)
- Andromeda (2)
- Astrophysics (2)
- Cnidaria (2)
- Cosmology (2)
- Discs (2)

Found 3 results. < 1 >

Sort by: Best match asc. View

June 1, 2001 (v1) Thesis Open Access

### Globular Clusters in the Andromeda Galaxy

Barmby, Pauline;

Globular clusters are among the oldest stellar systems, and the properties of their simple stellar populations provide valuable clues about galaxy formation. Local Group globular clusters form a bridge between Galactic and extragalactic globular clusters. The globular clusters of the Andromeda galaxy

Uploaded on May 18, 2016

View

May 16, 2011 (v1) Thesis Open Access

### Hydroxyl Masers from Andromeda to the Peak of Cosmic Star Formation

Willett, Kyle;

OH masers are naturally-occurring phenomena powered by stimulated emission, existing in a variety of astrophysical environments. The presence of powerful OH megamasers (OHMs) is associated with merging galaxies and extreme star formation, while the high luminosities and narrow beams of masers make t

Uploaded on August 24, 2015

View

April 19, 2011 (v1) Thesis Open Access

### Tracer populations in the Local Group

Watkins, Laura L;

So often in astronomy, an object is not considered for its individual merits, but for what we may learn from its properties regarding some larger population. The existence of dark matter is a prime example of this; we cannot see it directly but we can infer its presence by noting its effects on the

Uploaded on June 7, 2016

< 1 >


“The central stake”

from bating ational

son & al. vardship”  
<https://doi.org/10.1038/sdata.2016.18>

# Data Management & Stewardship



- Good Data Management and Stewardship is the key that leads to:
  - Knowledge discovery and innovation
  - Data and knowledge integration and reuse by the community
- The problem is far beyond long term data storage, since it includes data annotation  Metadata!

Goal: 


- transparency
- reproducibility
- reusability

 of data holdings for 

- humans
- machines

## Dataset discovery & integration perspective

- Do the datasets I'm searching for exist?
- Where are they published?
- How do I start the search?
- What tools I use?
- How do I access them?
- What formats are available?
- Can them be used together with
  - My local dataset?
  - Other dataset from different repositories?
- Can all of this be automated?
- Do licenses apply?

- Intuitive sense of semantics
  - Ability to identify directly the context(s)
- Less prone to error in selecting the data
  - Caveat: also humans need metadata
- Not fit to scope, scale, speed  We need machines!
  - Big Data

# Machine-driven activities



- Must be able to face wide range of
  - Types
  - Formats
  - Protocols
- Must keep provenance records

## ***“Machine Actionability”***

- Requires datasets with detailed information to move through autonomous action steps
  - Identify object type
  - Determine usefulness interrogating metadata
  - Determine usability: license, accessibility...
  - Take appropriate action



*“FAIR principles provide steps-along-the-path towards machine-actionability”*

- 2 contexts
  - Data discovery: contextual metadata
  - Data access: digital object content
- 2 approaches
  - Data-type specific support
  - General purpose open technologies

Ultimate goal: in the growing data environment, through general purpose repositories, guide the machines in finding and using datasets they have never seen before.

# Before going to details...



- A Data Object is defined as
  - An Identifiable Data Item
  - Data elements + Metadata + Identifier
- The term (meta)data indicates that the principle
  - is true for Metadata as well as Data Elements in the Data Object
  - can be independently implemented for either one
- FAIR principles
  - Related, Independent, Separable
  - To lower the entry barrier
  - To allow incremental implementation
    - But acting as guides, i.e. before implementation
- Good data management and stewardship is not a goal in itself, but rather a pre-condition supporting knowledge discovery and innovation

# The FAIR Guiding Principles



- To be Findable:
  - F1. (meta)data are assigned a globally unique and persistent identifier
  - F2. data are described with rich metadata
  - F3. metadata clearly and explicitly include the identifier of the data it describes
  - F4. (meta)data are registered or indexed in a searchable resource
- To be Accessible:
  - A1. (meta)data are retrievable by their identifier using a standardized communications protocol
    - A1.1 the protocol is open, free, and universally implementable
    - A1.2 the protocol allows for an authentication and authorization procedure, where necessary
  - A2. metadata are accessible, even when the data are no longer available
- To be Interoperable:
  - I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
  - I2. (meta)data use vocabularies that follow FAIR principles
  - I3. (meta)data include qualified references to other (meta)data
- To be Reusable:
  - R1. meta(data) are richly described with a plurality of accurate and relevant attributes
    - R1.1. (meta)data are released with a clear and accessible data usage license
    - R1.2. (meta)data are associated with detailed provenance
    - R1.3. (meta)data meet domain-relevant community standards

# Findable (1)



*F1. (meta)data are assigned a globally unique and persistent identifier*

- Data & metadata persistence
  - Actual requirement to publishers to state what is persisted
- Global uniqueness of identifiers
  - An handle to attach to each atomic concept
  - The identifier (URI, DOI, Handle.net, ...) uniquely points to the concept
- (meta)data are required an identifier being the “atomic” concepts that bring information

URI = scheme ":" hier-part [ "?" query ] [ "#" fragment ]  
<ivo://ia2.inaf.it/hosted/vipers/ssap>

The DOI syntax shall be made up of a DOI prefix and a DOI suffix separated by a forward slash.

[10.1038/issn.1476-4687](https://doi.org/10.1038/issn.1476-4687)

## *F2. data are described with rich metadata*

- Metadata are described throughout the principles, especially at “Re-usable” level
- Data are required proper description to allow actionability
  - Discovery context (findable)
- Metadata richness is required to
  - Distinguish among objects
  - Filter datasets using appropriate information

# Findable (3)



*F3. metadata clearly and explicitly include the identifier of the data it describes*

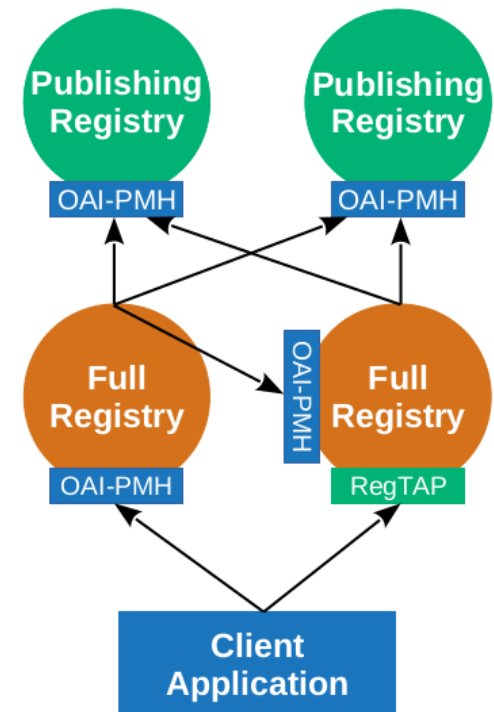
- Allow for data/metadata separation and reference
- Identifier become connection points in the ecosystem of data, metadata, services, tools that can build up a virtual research environment (VRE)

# Findable (4)



*F4. (meta)data are registered or indexed in a searchable resource*

- (meta)data are “collected” in catalogues or repositories
- Metadata allow for filtering (search) of (meta)data resources within the catalogues
- Catalogues themselves should be indexed and findable
  - Web search engine are brute force metadata repositories in a sense
  - One needs a bootstrapping mechanism



*A1. (meta)data are retrievable by their identifier using a standardized communications protocol*

*A1.1 the protocol is open, free, and universally implementable*

*A1.2 the protocol allows for an authentication and authorization procedure, where necessary*

- (meta)data are “resolvable” through the identifier
  - A mechanism exists that brings from the identifier to the content
    - URLs do exactly these, through DNS and web servers, HTTP being the protocol
- Protocol “open-ness” prevents barriers
- A&A is used in cases where sensible/restricted (meta)data is published



*A2. metadata are accessible, even when the data are no longer available*

- Persistence solution
  - Implemented also at PIDs repositories
- Metadata are considered as valuable as the data themselves

*I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.*

- Metadata is not simply free-text description
  - Machine-driven requirement
- (meta)data formats should be “open”
- All the information has be made available with machine-actionability in mind
  - web binary dumps are not FAIR
- Common (accessible, broadly accepted and shared) languages enable machine interoperation

## *12. (meta)data use vocabularies that follow FAIR principles*

- Vocabularies have to be FAIR data
  - Recursive requirement
- Vocabulary terms should be identified
  - The “context” should be qualified itself
- They have to be “parse-enabled”
- They have to use a common language
- Allow concepts to be re-used and understood outside their original context

*13. (meta)data include qualified references to other (meta)data*

- Interoperability requires actionable connections among (meta)data
- Qualified references are actually unique identifiers

*R1. meta(data) are richly described with a plurality of accurate and relevant attributes*

- Some points to take into consideration (non-exhaustive list):
  - Describe the scope of your data: for what purpose was it generated/collected?
  - Mention any particularities or limitations about the data that other users should be aware of.
  - Specify the date of generation/collection of the data, the lab conditions, who prepared the data, the parameter settings, the name and version of the software used.
  - Is it raw or processed data?
  - Ensure that all variable names are explained or self-explanatory (i.e., defined in the research field's controlled vocabulary).
  - Clearly specify and document the version of the archived and/or reused data. Metadata richness (see Find-ability)

*R1.1. (meta)data are released with a clear and accessible data usage license*

- Commonly used licenses
  - MIT
  - Creative Commons
  - GNU GPLv3
- Choose an Open Source Licence  
<https://choosealicense.com>

## *R1.2. (meta)data are associated with detailed provenance*

- As an example check this image  
[https://commons.wikimedia.org/wiki/File:Sampling\\_coral\\_microbiome\\_\(27146437650\).jpg](https://commons.wikimedia.org/wiki/File:Sampling_coral_microbiome_(27146437650).jpg)
- It includes:
  - authorship details
  - uses the Creative Commons Attribution Share Alike license
  - indicates exactly how the data author wishes to be cited

# Re-usable (4)

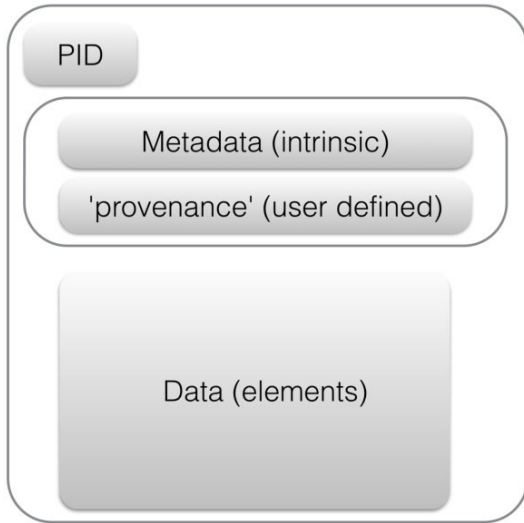


## *R1.3. (meta)data meet domain-relevant community standards*

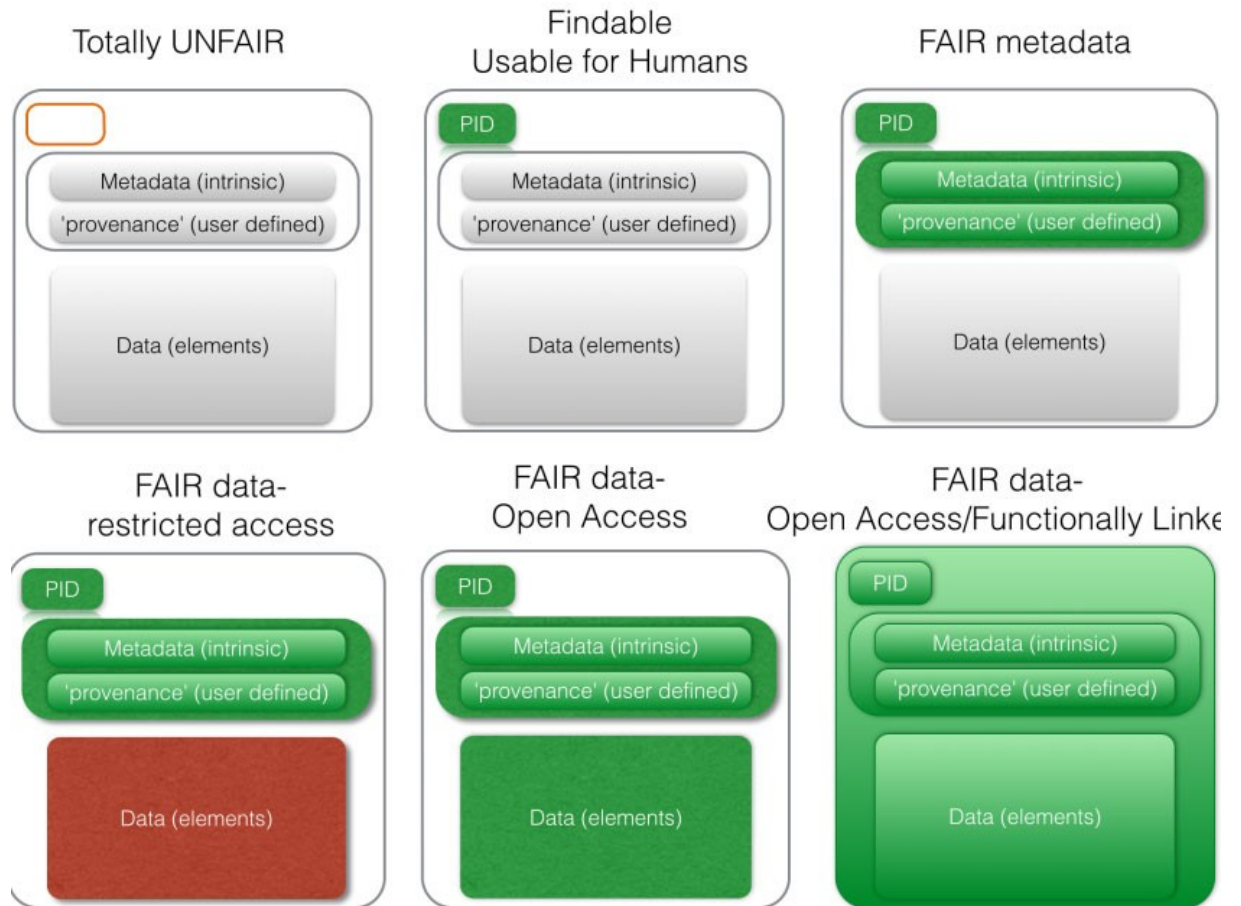
- List of some community standards
  - <http://schema.datacite.org/> [for general purpose, not domain-specific]
  - <http://dublincore.org/specifications/> [for general purpose, not domain-specific]
  - <https://www.ncbi.nlm.nih.gov/geo/info/MIAME.html> [microarrays]
  - <http://cds.u-strasbg.fr/doc/catstd.htx> [astrophysics]
  - <https://www.iso.org/standard/53798.html> [geographic information and services]
  - <http://cfconventions.org/> [climate and forecast]
  - <http://www.iucr.org/resources/cif> [crystallographic information]
  - <http://www.nexusformat.org/> [neutron, x-ray, and muon experiment data]
  - <http://www.ddialliance.org/Specification> [social, behavioral, and economic sciences]
  - <https://sdmx.org/> [statistical data]
  - <https://knb.ecoinformatics.org/#tools/eml> [ecology]



# Digital Data FAIRness



## Data as increasingly FAIR Digital Objects



# Example: Dataverse



- Dataverse is an open-source data repository software installed in dozens of institutions globally to support public community repositories or institutional research data repositories.
- Harvard Dataverse, with more than 60,000 datasets, is the largest of the current Dataverse repositories, and is open to all researchers from all research fields.
- Dataverse makes the Digital Object Identifier (DOI), or other persistent identifiers (Handles), public when the dataset is published ('F').
  - DOI resolves to a landing page, providing access to metadata, data files, dataset terms, waivers or licenses, and version information, all of which is indexed and searchable ('F', 'A', and 'R').
  - Dataverse generates a formal citation for each deposit, following defined standard ('R')
- Deposits include metadata, data files, and any complementary files (such as documentation or code) needed to understand the data and analysis ('R').
- Metadata is always public, even if the data are restricted or removed for privacy issues ('F', 'A'). This metadata is offered at three levels, extensively supporting the 'I' and 'R' FAIR principles: 1) data citation metadata, which maps to DataCite schema or Dublin Core Terms, 2) domain-specific metadata, which when possible maps to metadata standards used within a scientific domain, and 3) file-level metadata, which can be deep and extensive for tabular data files (including column-level metadata).
- Finally, Dataverse provides public machine-accessible interfaces to search the data, access the metadata and download the data files, using a token to grant access when data files are restricted ('A').



Open source research data repository software

# Example: FAIRDOM



- FAIRDOM integrates several platforms to produce a FAIR data and model management facility for Systems Biology.
  - SEEK is a web-based resource for sharing heterogeneous scientific research datasets, models or simulations, processes and research outcomes.
  - Rightfield is an open-source tool for adding ontology term selection to spreadsheets.
  - JWS Online is a systems biology tool for the construction, modification and simulation of kinetic models
  - OpenBIS is an open, distributed system for managing biological information of data workflows.
- Individual research assets (or aggregates of data and models) are identified with unique and persistent HTTP URLs, which can be registered with DOIs for publication ('F', 'A').
- Assets can be accessed over the Web in a variety of formats appropriate for individuals and/or their computers (RDF, XML) ('I', 'A').
- Research assets are annotated with rich metadata, using community standards, formats and ontologies ('I').
- The metadata is stored as RDF to enable interoperability and assets can be downloaded for reuse ('R').
- See the video “What is data management?” on <https://fair-dom.org>



# Example: Open PHACTS



- Open PHACTS is a data integration platform for information pertaining to drug discovery.
- Access to the platform is mediated through a machine-accessible interface which provides multiple representations that are both human (HTML) and machine readable (RDF, JSON, XML, CSV, etc), providing the 'A' facet of FAIRness.
- The interface allows multiple URLs to be used to access information about a particular entity through a mappings service ('F' and 'A').
- Thus, a user can provide a ChEMBL URL to retrieve information sourced from, for example, Chempider or DrugBank. Each call provides a canonical URL in its response ('A' and 'I').
- All data sources used are described using standardized dataset descriptions, following the global standard, with rich provenance ('R' and 'I').
- All interface features are described using RDF following the Linked Data API specification ('A').
- Finally, a majority of the datasets are described using community agreed upon ontologies ('I').



---

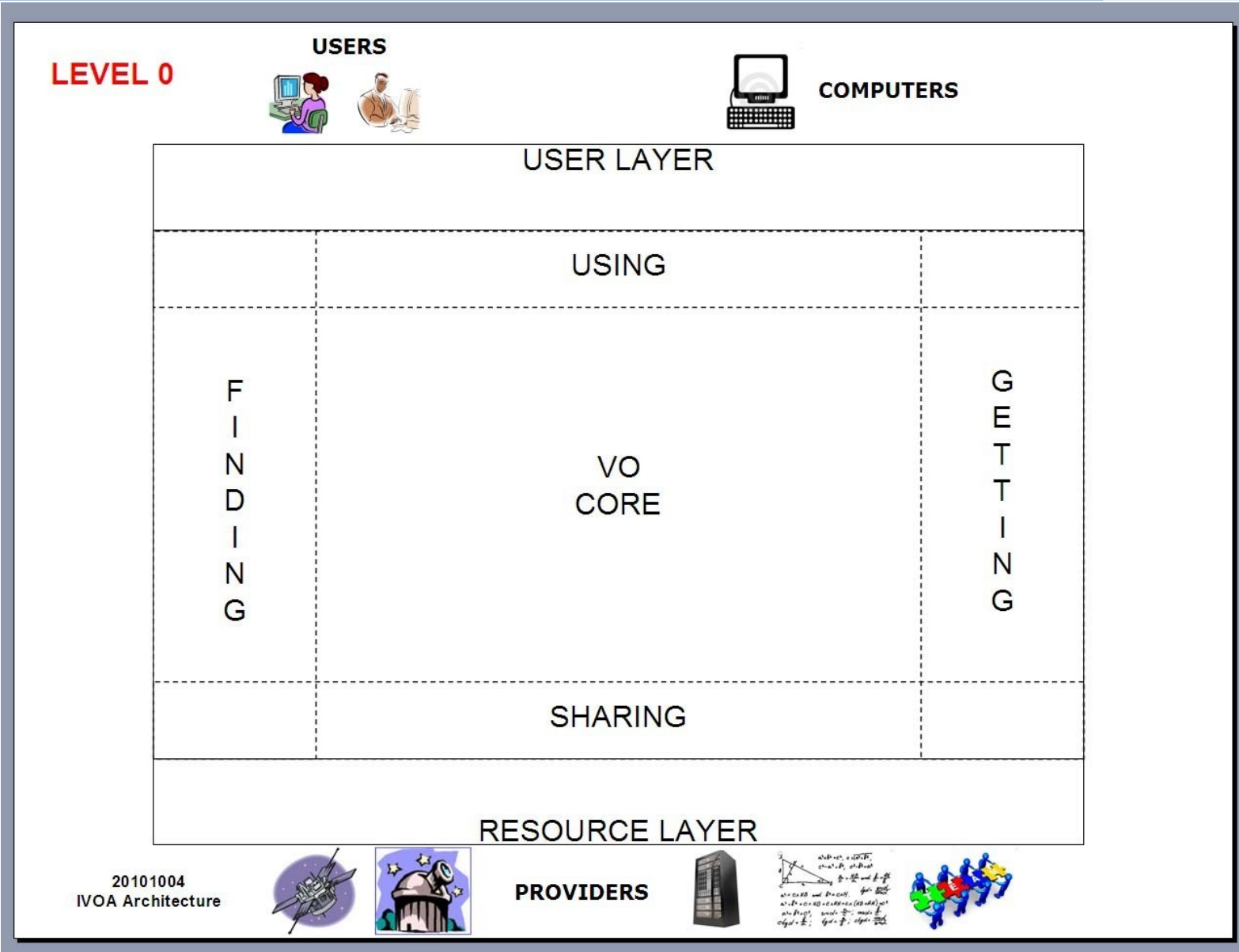
Bringing together **pharmacological data resources** in an integrated, interoperable infrastructure

# Example: IVOA (1)

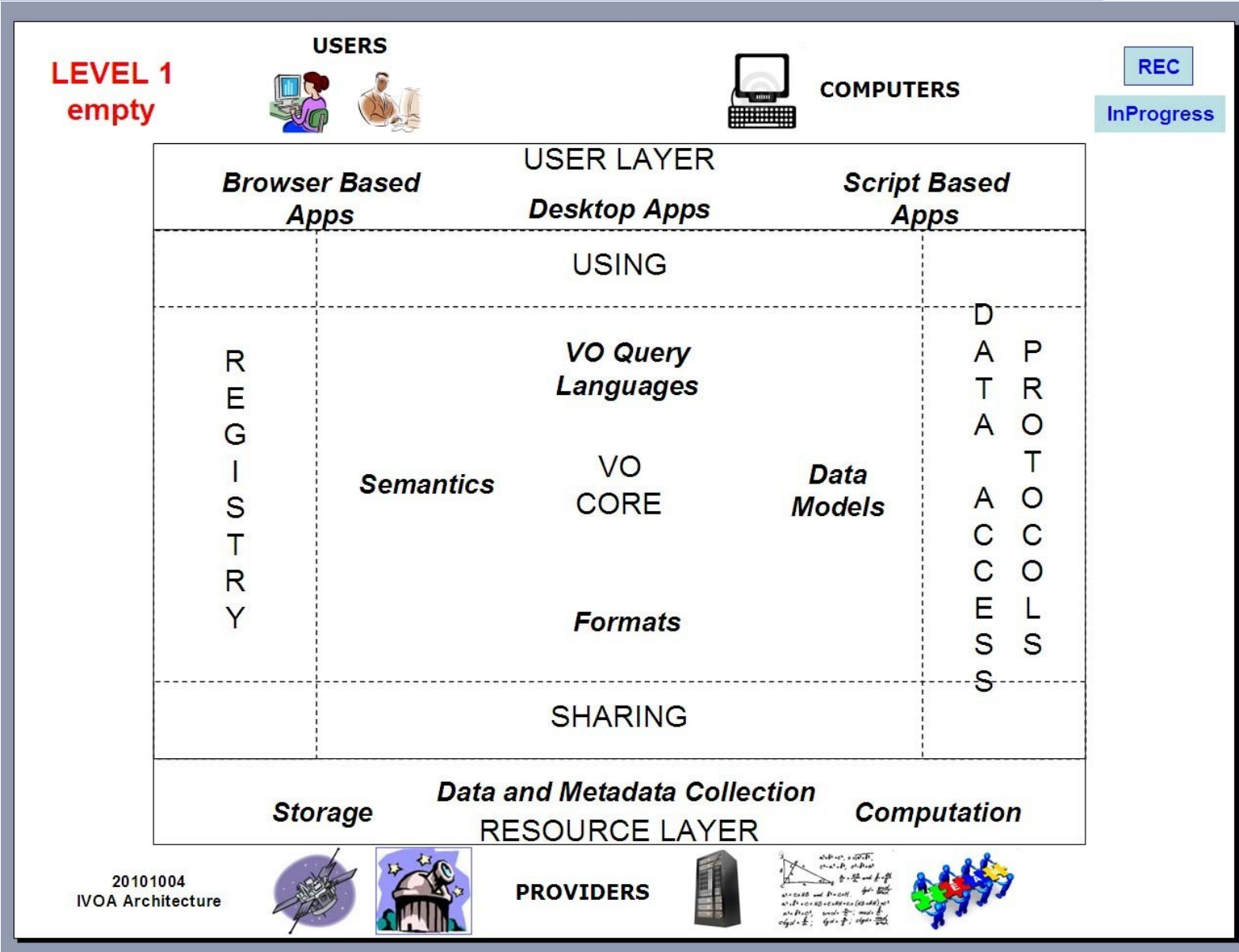


- The Virtual Observatory (VO) is the vision that astronomical datasets and other resources should work as a seamless whole.
- Many projects and data centres worldwide are working towards this goal.
- The International Virtual Observatory Alliance (IVOA) is an organisation that debates and agrees the technical standards that are needed to make the VO possible.
- IVOA also acts as a focus for VO aspirations, a framework for discussing and sharing VO ideas and technology, and body for promoting and publicising the VO.

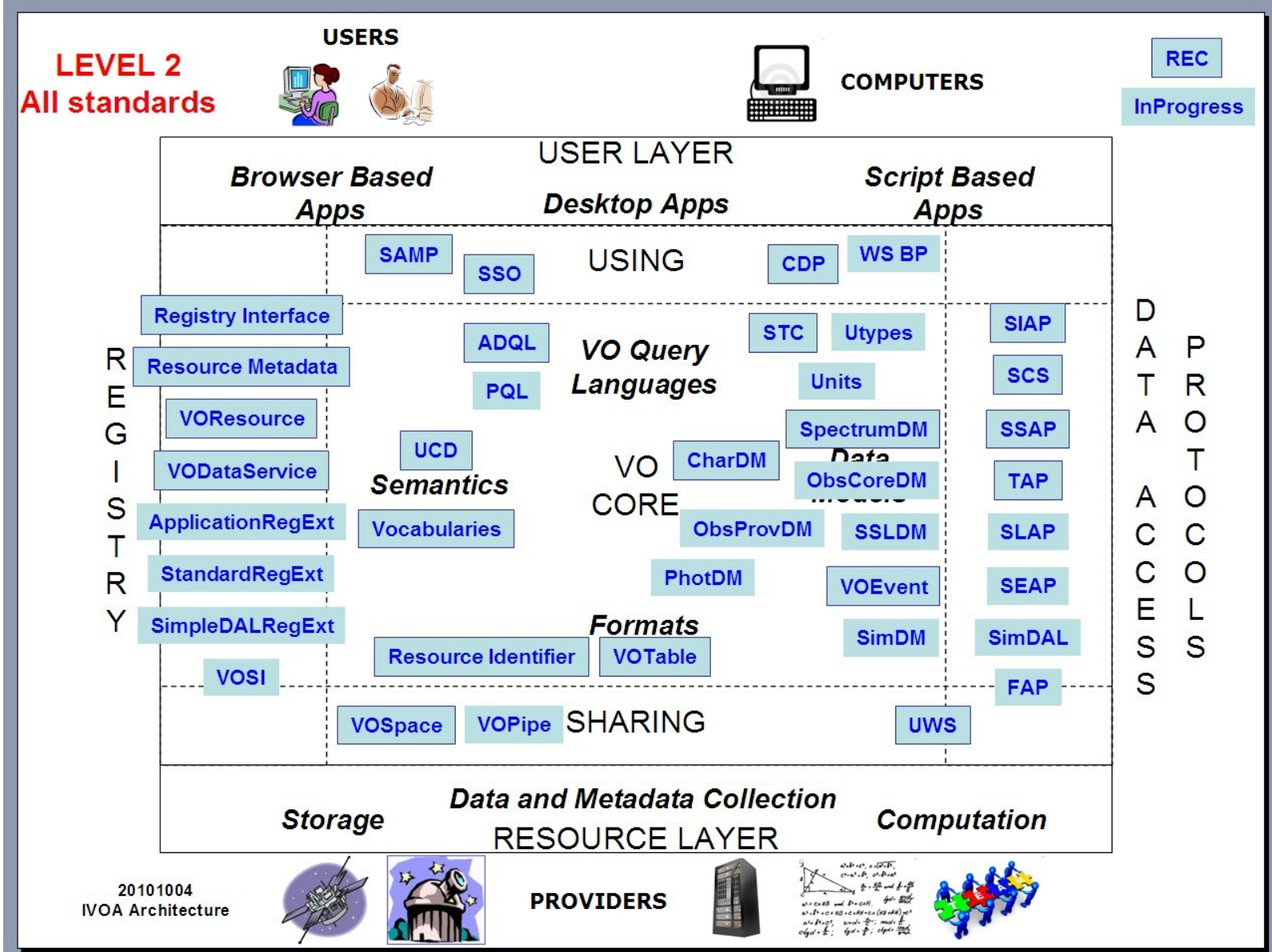
# Example: IVOA (2)



# Example: IVOA (3)

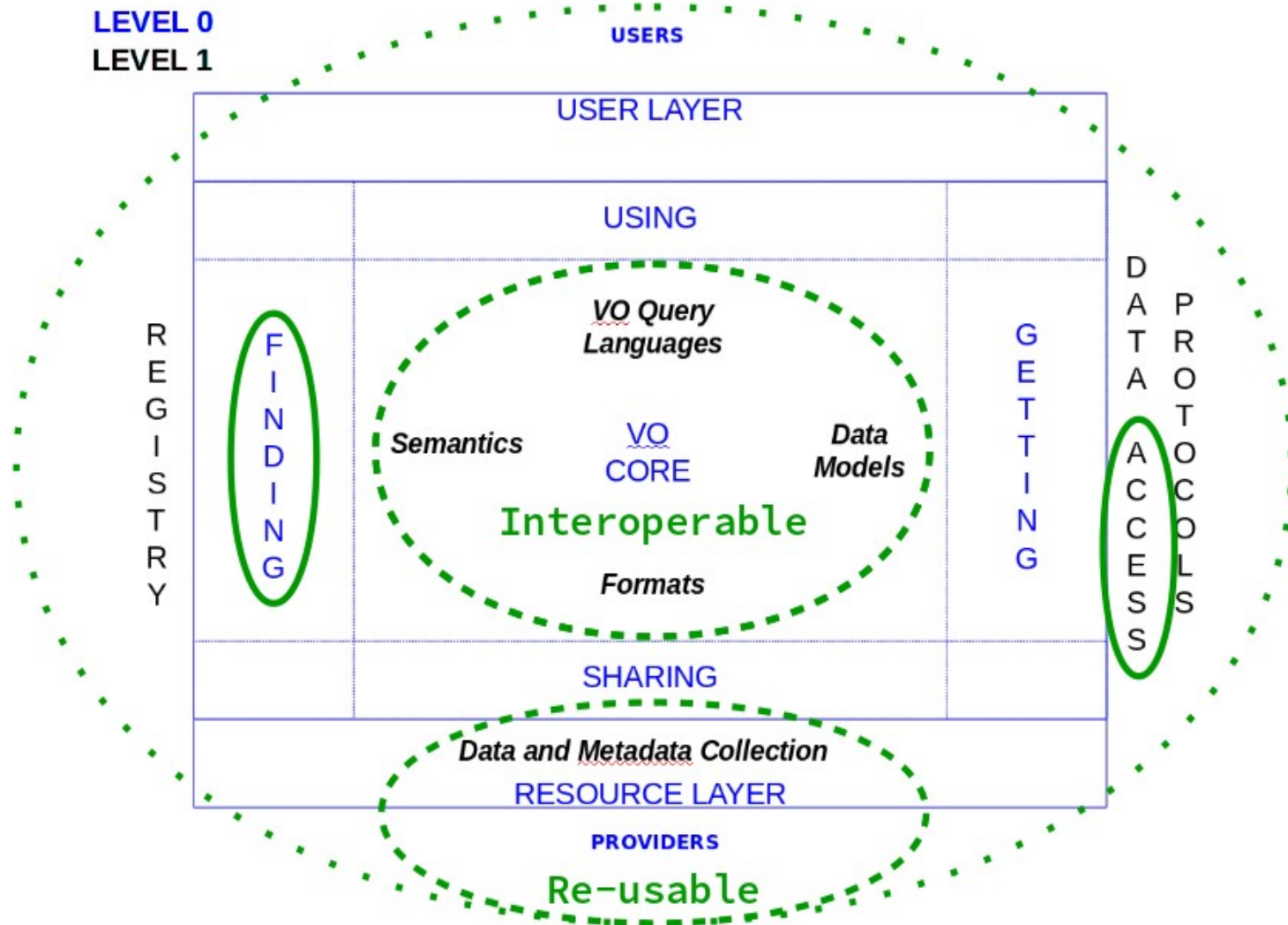


# Example: IVOA (4)





# Example: IVOA (5)



# References



- DOI: 10.1038/sdata.2016.18
- <https://zenodo.org>
- <http://simbad.u-strasbg.fr/simbad/>
- <https://www.go-fair.org>
- <https://www.force11.org/fairprinciples>
- <https://www.doi.org>
- <https://dataverse.org>
- <https://fair-dom.org>
- <https://www.openphacts.org>
- <http://ivoa.net>