

LSDMA Topical Meeting 6

Data Intensive Computing and Applications

Lambda

Integration of Data Repository System and Data Intensive Computing



T. Jejkal, S. Chandna, R. Dapp, V. Hartmann, A. Prabhune, F. Rindone,
R. Stotzka, D. Tonne, A. Vondrous, X. Yang



Repository System

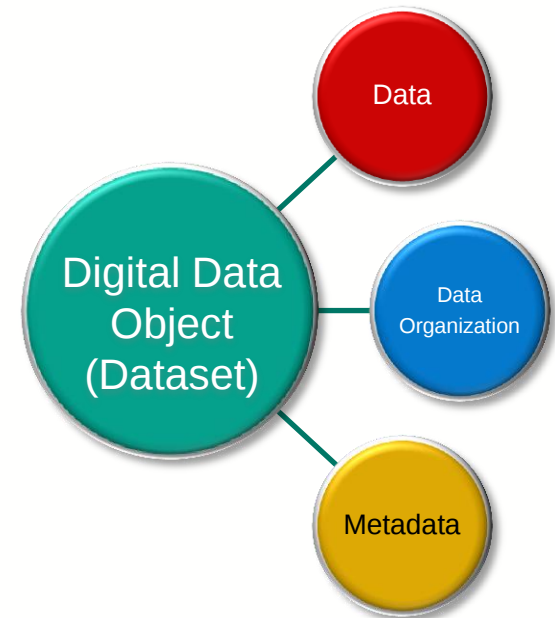
Repository

Managed location/destination/directory/bucket where **digital data objects** are

- Registered
- Permanently stored
- Made accessible and retrievable
- Curated

Digital data object (dataset) consists of

- Data
- Description for re-use



What happens to Digital Data Objects?

Access and Retrieval

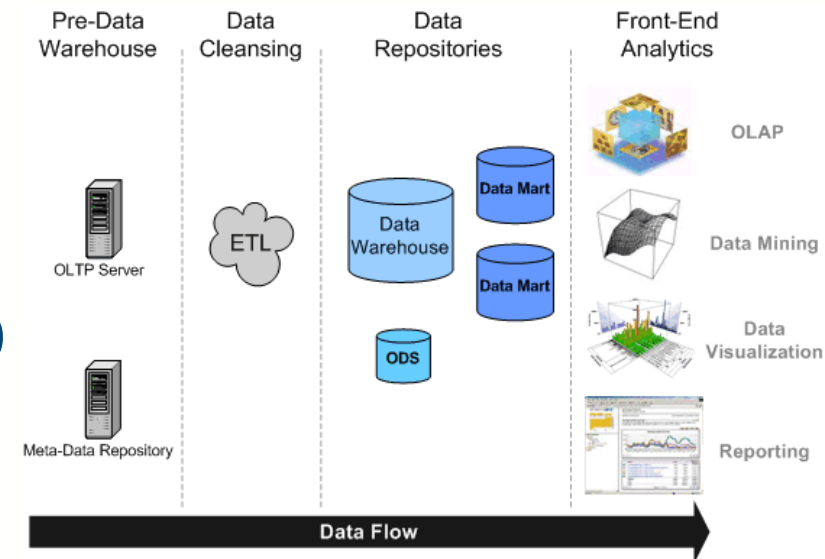
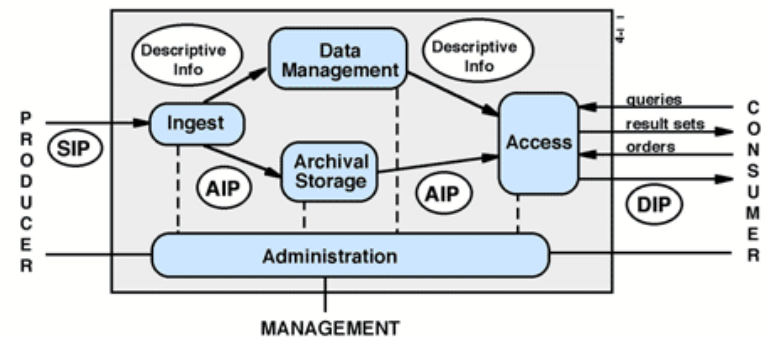
Query, download, sharing

Preservation

Bit- and content preservation, curation

Analytics

Data visualization and -mining,
Online Analytical Processing (OLAP),
requires “Extract, Transform, Load” (ETL)



http://www.dwreview.com/Images/DataWarehouse_Overview.gif
http://nssdc.gsfc.nasa.gov/nssdc_news/dec00/oais_fig3.gif

What is different for Research Data?

It starts with raw data

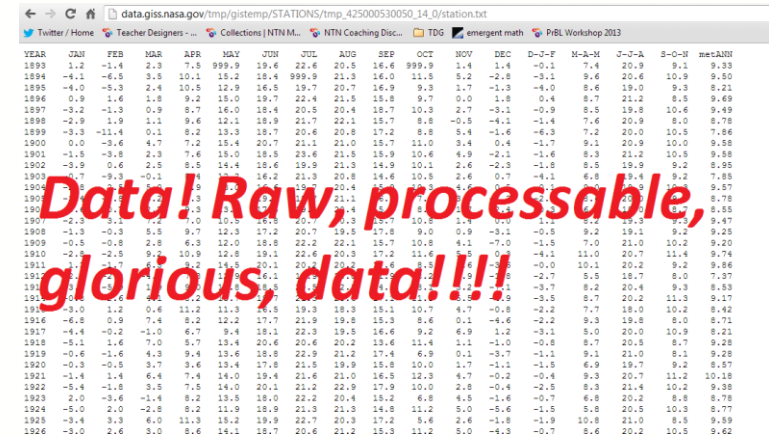
- Raw data is the most valuable data
- Processing is reproducible, DAQ not

Research data can be huge

- Typical datasets in range of GB/TB
- Local/on-the-fly processing not applicable
- Even management tasks get time consuming

Trying to understand research data can be challenging

- Complex processing chain of novel algorithms
- Many datasets (re-)processed with same algorithms
- Provenance information to tweak/compare algorithms and parameters



data.giss.nasa.gov/imp/gistemp/STATIONS/tmp_425000530050_14_0/station.txt

Twitter / Home Teacher Designers - ... Collections | NTN M... NTN Coaching Disc... TDG emergent math PiBL Workshop 2013

YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	D-J-F	M-A-M	J-J-A	S-O-N	maxJUN
1893	1.2	-1.4	2.3	7.8	999.9	19.6	22.6	20.5	16.6	999.9	1.4	1.4	-0.1	7.4	20.9	9.1	9.33
1894	-4.1	-6.5	3.5	10.1	15.2	18.4	999.9	21.3	16.0	11.5	5.2	-2.8	-3.1	9.6	20.6	10.9	9.50
1895	-4.0	-5.3	2.4	10.5	12.9	16.5	19.7	20.7	16.9	9.3	1.7	-1.3	-4.0	8.6	19.0	9.3	9.21
1896	0.9	1.6	1.8	9.2	15.0	19.7	22.4	21.5	15.8	9.7	0.0	1.8	0.4	8.7	21.2	8.5	9.69
1897	-3.2	-1.3	0.9	8.7	16.0	18.4	20.5	20.4	18.7	10.3	2.7	-3.1	-0.9	8.5	19.8	10.6	9.49
1898	-2.9	-1.9	1.1	9.6	12.1	18.9	21.7	22.1	15.7	8.8	-0.5	-4.1	-1.4	7.6	20.9	8.0	8.78
1899	-3.3	-11.4	0.1	8.2	13.3	18.7	20.6	20.8	17.2	8.8	5.4	-1.6	-6.3	7.2	20.0	10.5	7.86
1900	0.0	-3.6	4.7	7.2	15.4	20.7	21.1	21.0	15.7	11.0	3.4	0.4	-1.7	9.1	20.9	10.0	9.58
1901	-1.5	-3.8	2.3	7.6	15.0	18.5	23.6	21.5	15.9	10.6	4.9	-2.1	-1.6	8.3	21.2	10.5	9.58
1902	-3.9	0.6	2.5	8.5	14.4	18.6	19.5	21.9	16.9	10.1	2.6	-2.3	-1.5	8.5	19.9	9.2	8.95
1903	-0.7	-9.3	-0.1	7.1	15.5	16.2	21.3	20.8	14.6	10.5	2.6	0.7	-4.1	6.8	19.4	9.2	7.85
1904	-1.3	-1.7	1.9	8.9	15.7	19.7	20.4	21.3	15.7	10.4	4.5	0.7	-1.1	8.2	20.9	9.2	9.57
1905	-1.3	-1.7	1.9	8.9	15.7	19.7	20.4	21.3	15.7	10.4	4.5	0.7	-1.1	8.2	20.9	9.2	9.57
1906	-1.3	-1.7	1.9	8.9	15.7	19.7	20.4	21.3	15.7	10.4	4.5	0.7	-1.1	8.2	20.9	9.2	9.57
1907	-2.3	-1.2	1.2	1.0	10.3	26.8	20.7	19.3	15.7	10.1	1.4	0.0	-1.1	8.2	19.3	9.2	9.47
1908	-1.3	-0.3	5.5	9.7	12.3	17.2	20.7	19.5	17.8	9.0	0.9	-3.1	-0.5	9.2	19.1	9.2	9.25
1909	-0.5	-0.5	2.8	6.3	12.0	18.8	22.2	22.1	15.7	10.8	4.1	-7.0	-1.5	7.0	21.0	10.2	9.20
1910	-2.1	-2.5	9.1	10.9	12.8	19.1	22.6	20.3	17.2	11.6	3.5	0.0	-4.1	11.0	20.7	11.4	9.74
1911	-1.1	-0.7	1.2	1.2	14.1	19.1	20.1	20.1	15.7	10.1	1.6	-0.8	-0.0	10.1	20.2	9.2	9.86
1912	-1.1	-0.7	1.2	1.2	14.1	19.1	20.1	20.1	15.7	10.1	1.6	-0.8	-0.0	10.1	20.2	9.2	9.86
1913	-1.1	-0.7	1.2	1.2	14.1	19.1	20.1	20.1	15.7	10.1	1.6	-0.8	-0.0	10.1	20.2	9.2	9.86
1914	-1.1	-0.7	1.2	1.2	14.1	19.1	20.1	20.1	15.7	10.1	1.6	-0.8	-0.0	10.1	20.2	9.2	9.86
1915	-1.1	-0.7	1.2	1.2	14.1	19.1	20.1	20.1	15.7	10.1	1.6	-0.8	-0.0	10.1	20.2	9.2	9.86
1916	-1.1	-0.7	1.2	1.2	14.1	19.1	20.1	20.1	15.7	10.1	1.6	-0.8	-0.0	10.1	20.2	9.2	9.86
1917	-4.4	-0.2	-1.0	6.7	9.4	18.1	22.3	19.5	16.6	9.2	6.9	1.2	-3.1	5.0	20.0	10.9	8.21
1918	-5.1	1.6	7.0	5.7	13.4	20.6	20.6	20.2	13.6	11.4	1.1	-1.0	-0.8	8.7	20.8	8.7	9.28
1919	-0.6	-1.6	4.3	9.4	13.6	18.8	22.9	21.2	17.4	6.9	0.1	-3.7	-1.1	9.1	21.0	8.1	9.28
1920	-0.3	-0.5	3.7	3.6	13.4	17.8	21.5	19.9	15.8	10.0	1.7	-1.1	-1.5	6.9	19.7	9.2	8.57
1921	-1.4	1.4	6.4	7.4	14.0	19.4	21.6	21.0	16.5	12.3	4.7	-0.2	-0.4	9.3	20.7	11.2	10.18
1922	-5.4	-1.8	3.5	7.5	14.0	20.1	21.2	22.9	17.9	10.0	2.8	-0.4	-0.5	8.3	21.4	10.2	9.38
1923	2.0	-3.6	-1.4	8.2	13.5	18.0	22.2	20.4	15.2	6.8	4.5	-1.6	-0.7	6.8	20.2	8.8	8.78
1924	-5.0	2.0	-2.8	8.2	11.9	18.9	21.3	21.3	14.8	11.2	5.0	-5.6	-1.5	5.8	20.5	10.3	8.77
1925	-3.4	3.3	6.0	11.3	15.2	19.9	22.7	20.3	17.2	5.6	2.6	-1.8	-1.9	10.8	21.0	8.5	9.59
1926	-3.0	2.6	3.0	8.6	14.1	18.7	20.6	21.2	15.3	11.2	5.0	-4.3	-0.7	8.6	20.2	10.5	9.62

<https://emergentmath.files.wordpress.com/2013/03/untitled2.png>

What is needed?



A research data repository system

- Support large scale research data
- Flexibility to be enhanced by DIC

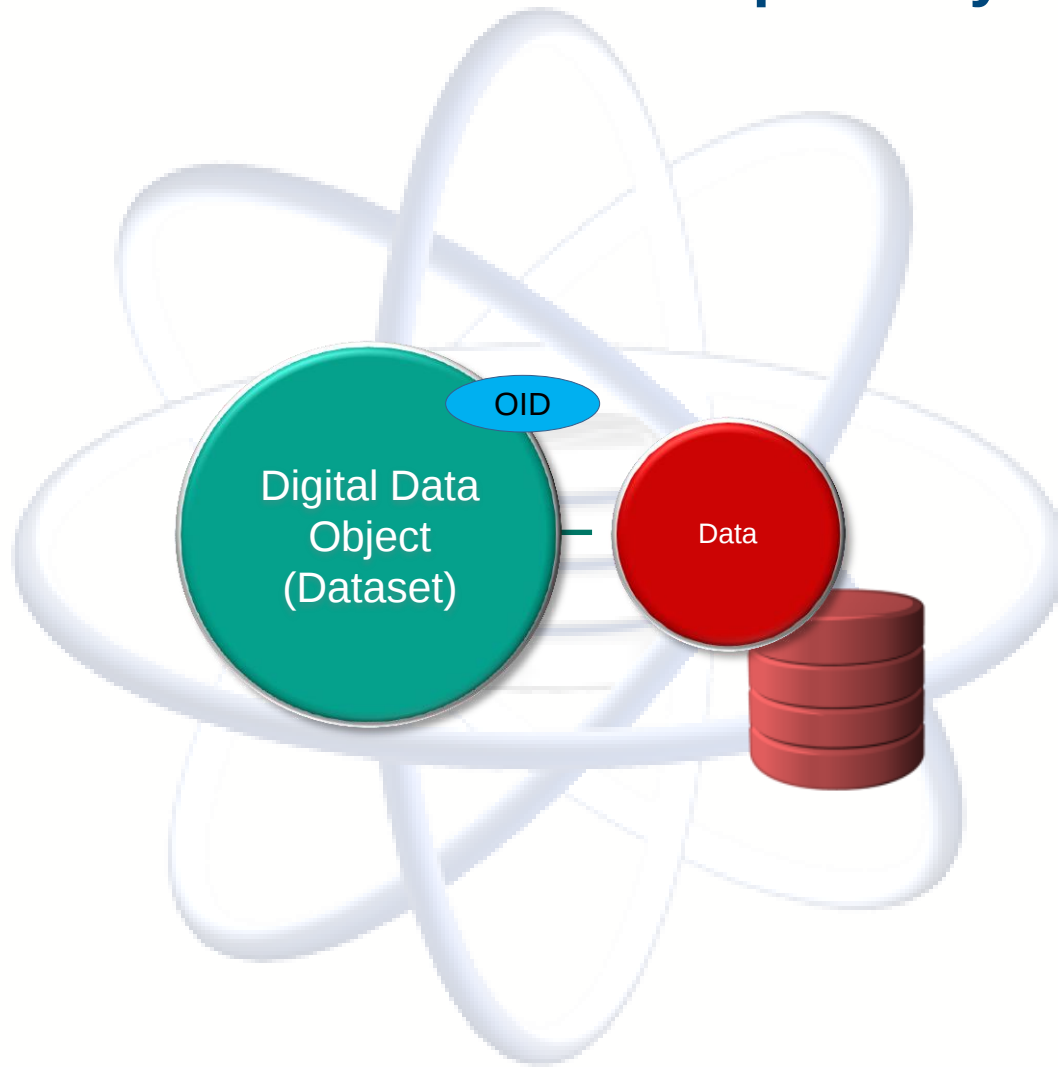
Description by metadata

- High-level integration and flexibility
- Allows reproducibility and provenance information tracking
- Support for comparing different configurations

Seamless integration

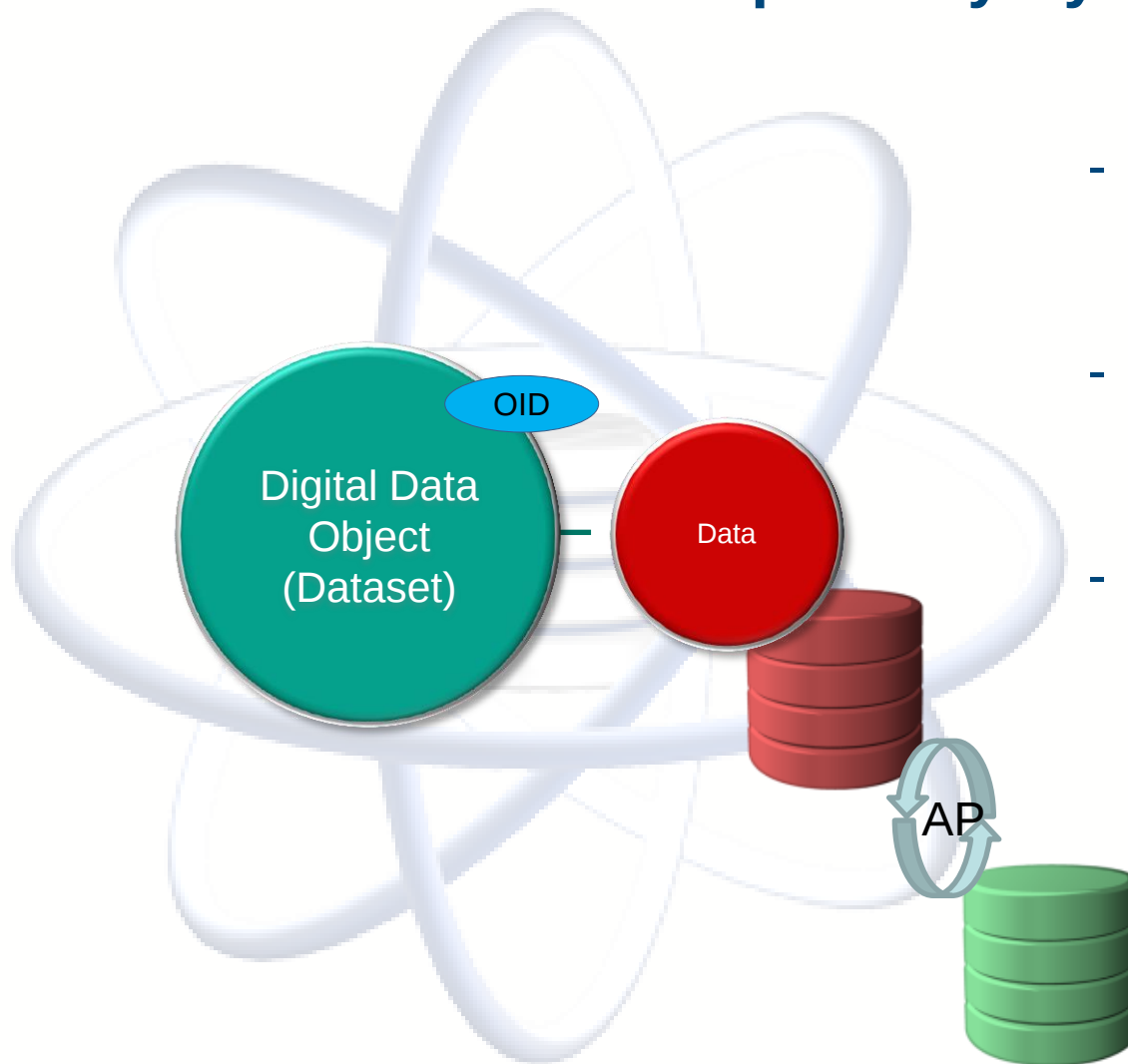
- Computing should integrate into repository workflows
- No manual data movement
- Should be usable for system- and user-applications

The Research Data Repository System

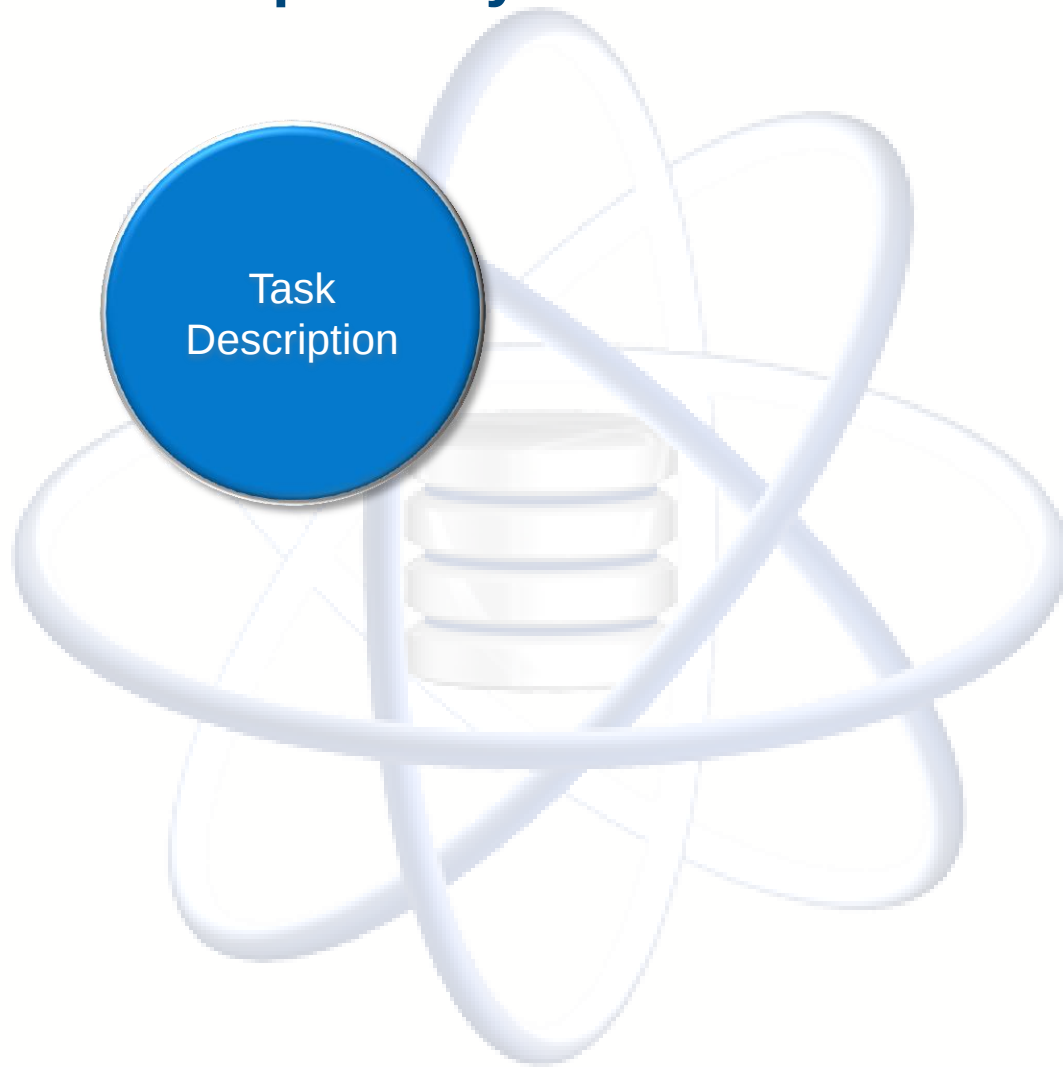


- Datasets registered in repository system identified by ObjectId
- Data available in repository storage accessible by repository system

The Research Data Repository System

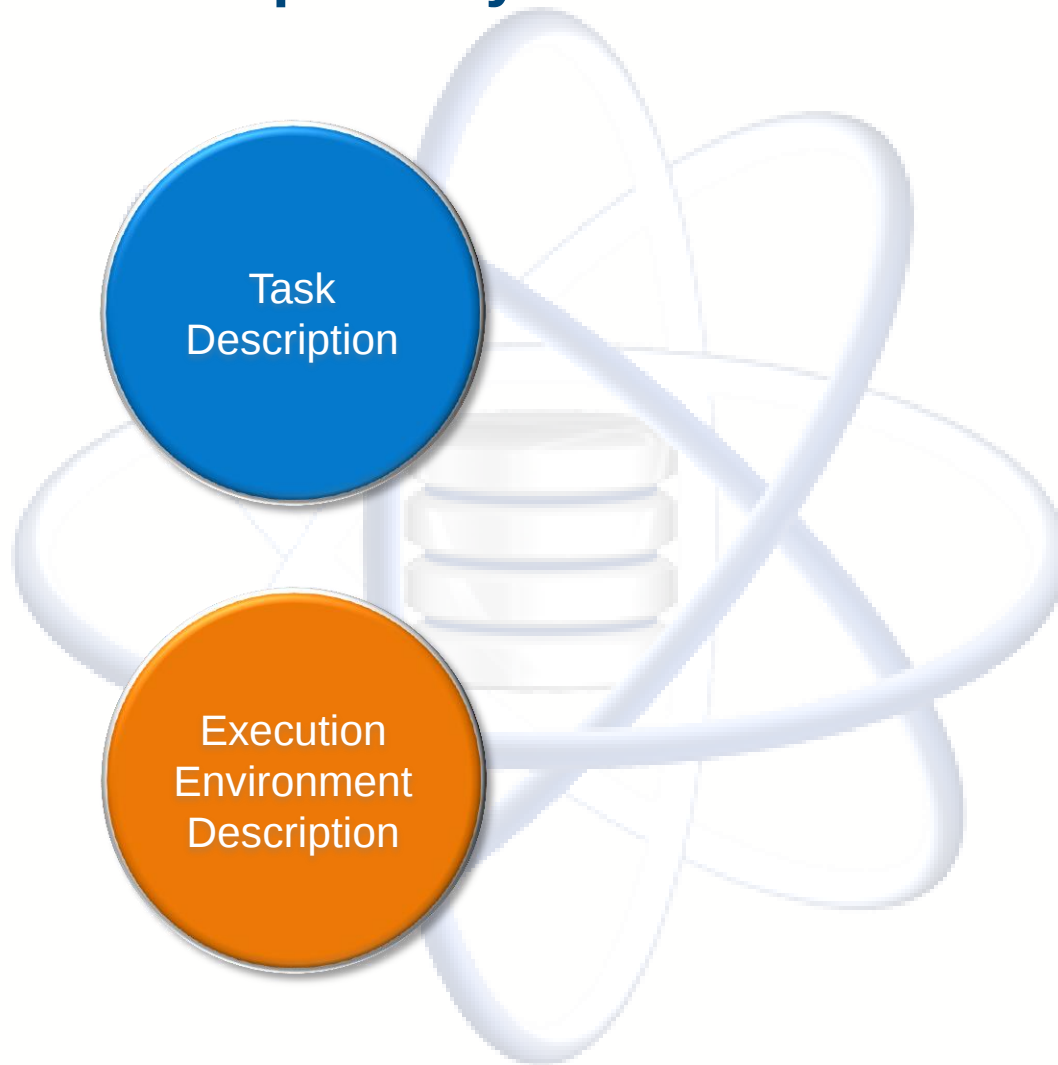


- Datasets registered in repository system identified by ObjectId
- Data available in repository storage accessible by repository system
- Configurable access points (AP) for data access/provisioning



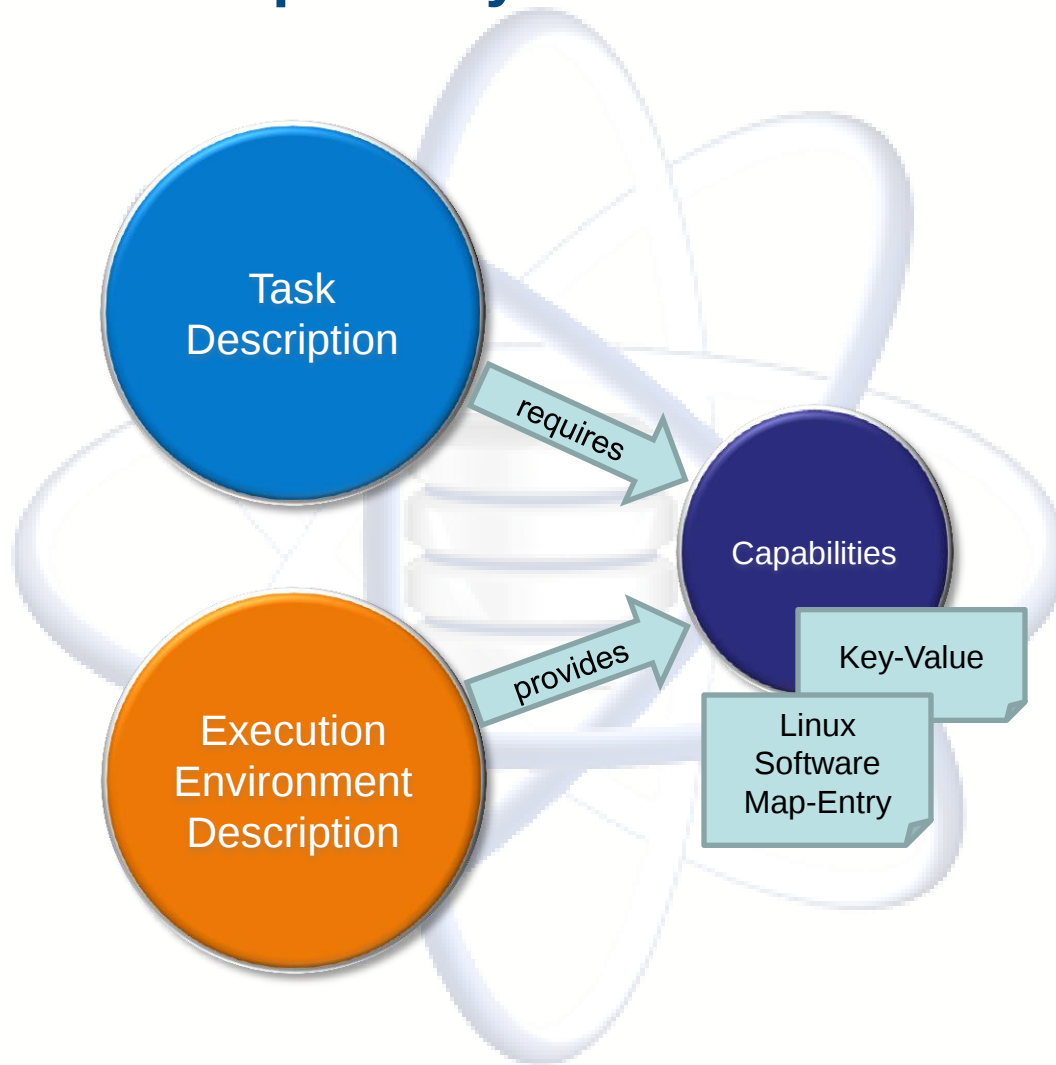
- Description of task including arguments, version and binary package

Description by Metadata



- Description of task including arguments, version and binary package
- Basic description of execution environment, e.g. repository AccessPoint, custom properties and handler implementation

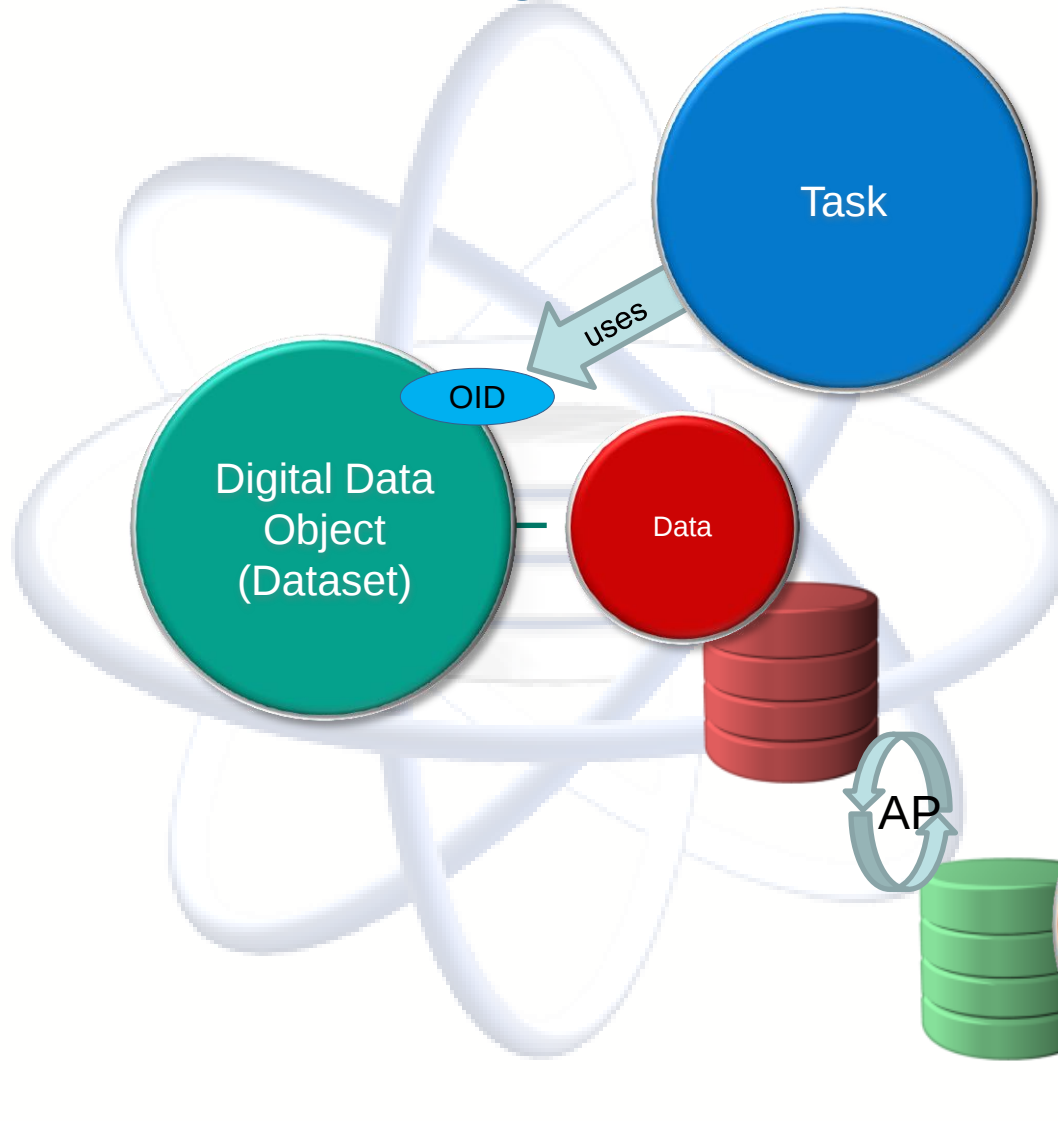
Description by Metadata



- Description of task including arguments, version and binary package
- Basic description of execution environment, e.g. usable AccessPoint, custom properties and handler implementation
- Capabilities to allow basic decision on task-environment assignment

→ **Application testing still necessary!**

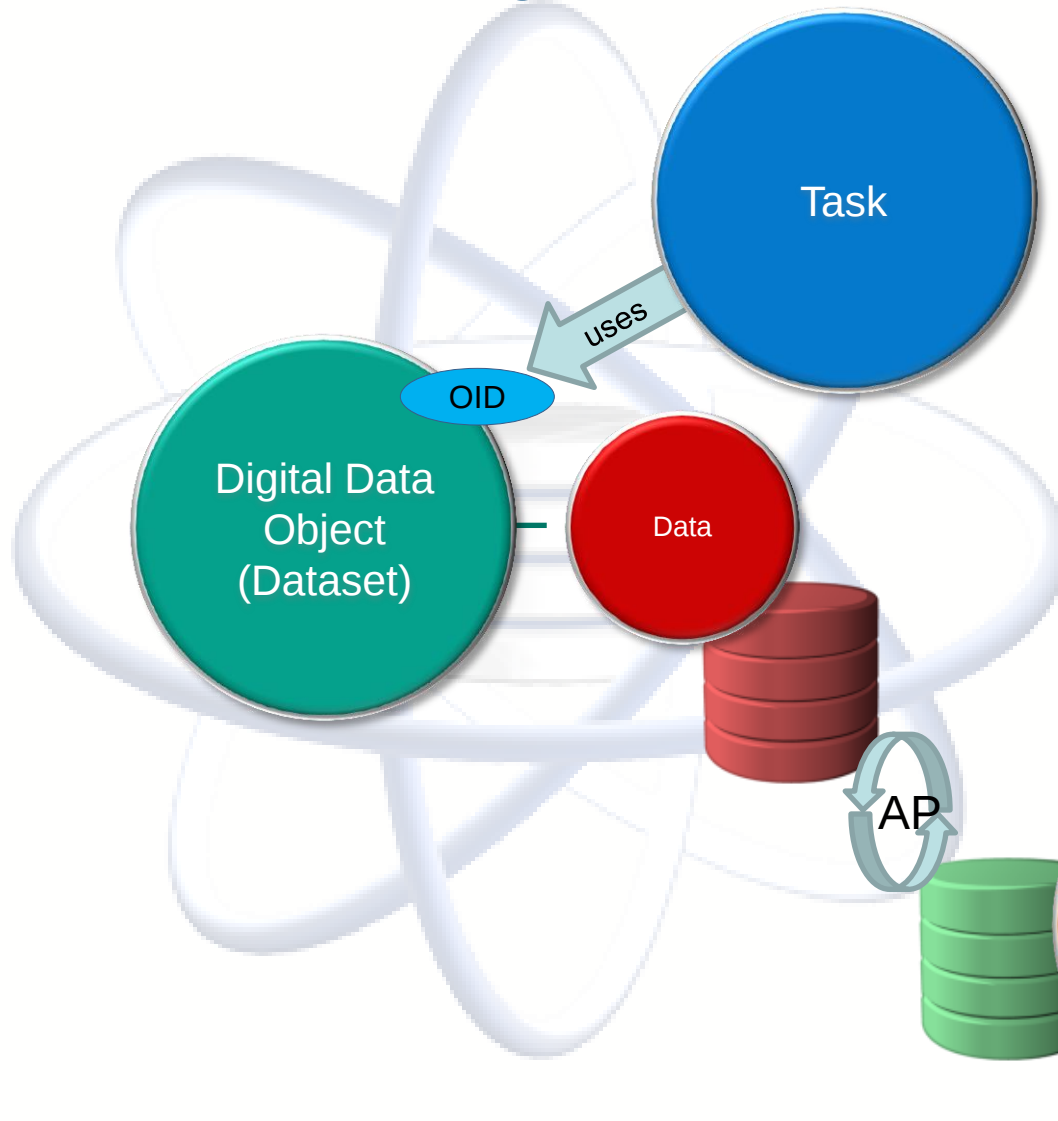
Seamless Integration



- Task instance linked to one or more OIDs
- Task execution component of repository system takes care of capability matching, data staging, submission, monitoring and result registration



Seamless Integration



- Task instance linked to one or more OIDs
- Task execution component of repository system takes care of capability matching, data staging, submission, monitoring and result registration

- First integration of DIC capabilities into repository system
- Metadata description of tasks including arguments, execution environment and capabilities
- Flexible execution environment handler interface
- Used applications must be well tested (focus on recurring tasks)
- Currently first evaluation, afterwards integration of Apache Flink
- Support for provenance standards OPM and PROV



IG Research Data Provenance