A Grid Environment for Data Integration of Scientific Databases

Hideo Matsuda[1],
Susumu Date[1], Shinji Shimojo[1]
(1 Osaka University)
Kentaro Wakatsuki[2], Takehiro Furudate[2]
Gen Kawamura[1,3], Yoshiyuki Kido[1,4]
(2 Hitachi Software, 3 Aztec System, 4 Mitsui Knowledge Industry)

Outline of Talk

- Genome Annotation Pipeline
- Introduction to Japan NAREGI Project
- NAREGI Data Grid
- Scientific DBs (focus: Lifescience DBs)
- Data Integration of Lifescience DBs

Genome Annotation Projects

- Inspired by the Drosophila Genome Annotation Jumboree.
- Several full-length cDNA (FLcDNA = clone of gene transcript) annotation projects have been organized in Japan (mouse, human and rice genomes).
- In the projects, FANTOM (*Functional ANnoTation Of Mouse*) is the first annotation project for genome FLcDNA sequences.
- Organized by RIKEN GSC.
- 1st FANTOM Meeting (FANTOM1) at RIKEN Tsukuba Institute
  - Aug. 28 ~ Sep. 8, 2000.
  - 21K FLcDNA seqs.
  - >60 researchers (>6 countries)
  - *Nature*, 409:685-690, 2001.
  - FANTOM DB:

  http:// fantom.gsc.riken.jp/
- FANTOM2 (2002) 60K seqs at RIKEN Yokohama Institute

  Nature, 420:563-573, 2002
- FANTOM3 (2004) 103K seqs at RIKEN Yokohama Institure

  Science, 309:1559-1563, 2005

Annotation Pipeline at FANTOM

(Kasukawa et al., *Genome Res.* 13:1542, 2003)

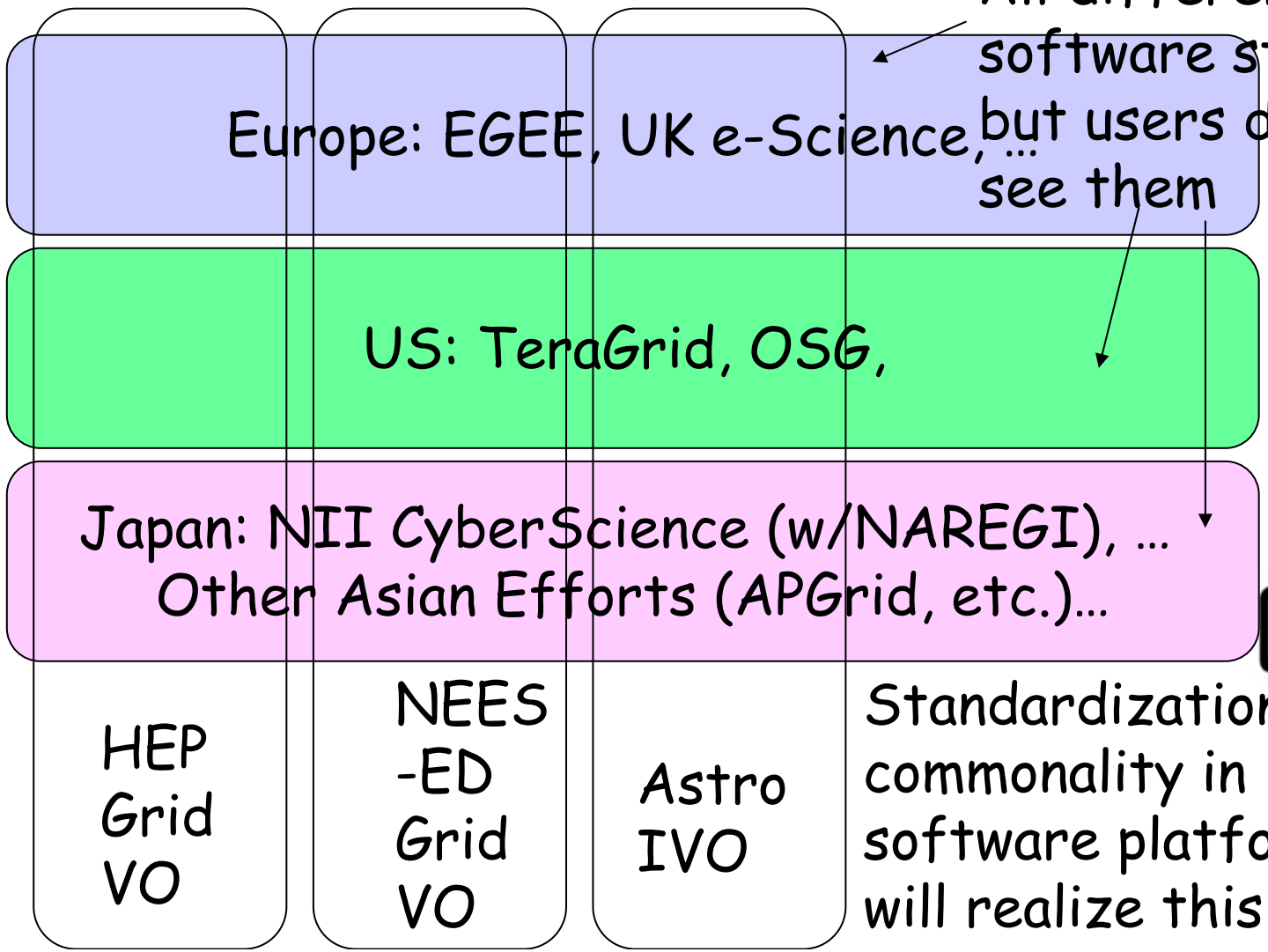
1. Clustering of FLcDNA seqs w/ FLAST(DDS)+ CAP3 and ClustalW
2. Masking repetitive elements w/ Repbase (RepeatMasker).
3. ORF (Coding Sequence) Prediction w/ RIKEN Decoder.
4. Sequence homology search (FASTY and BLASTX) against NCBI nr., SwissProt/TrEMBL nr.
5. Motif / Domain search w/ Pfam(estwise), InterPro (InterProScan), etc.
6. Mapping to mouse & human Genome (NCBI build) w/ RIKEN Genomapper.
7. Assignment of Gene Ontology (GO) terms.

blue: tools, red: data or databases

The Ideal World: Ubiquitous VO & user management for international e-Science

By courtesy of Prof. Matsuoka (Tokyo Inst. Tech.)

Grid Regional Infrastructural Efforts
Collaborative talks on PMA, etc.

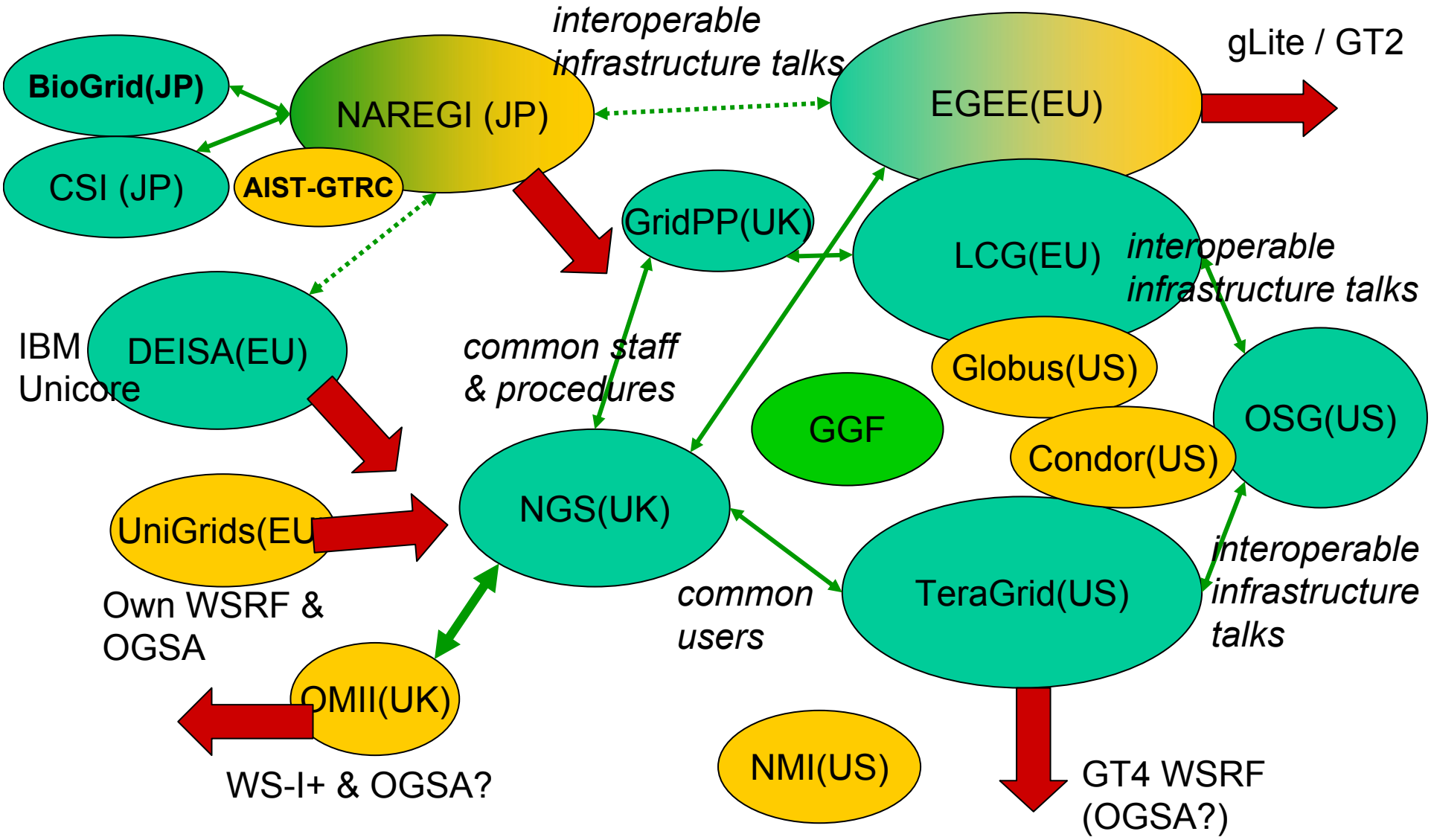


All different software stacks but users don't see them

Standardization, commonality in software platforms will realize this

The Reality: Convergence/Divergence of Project Forces

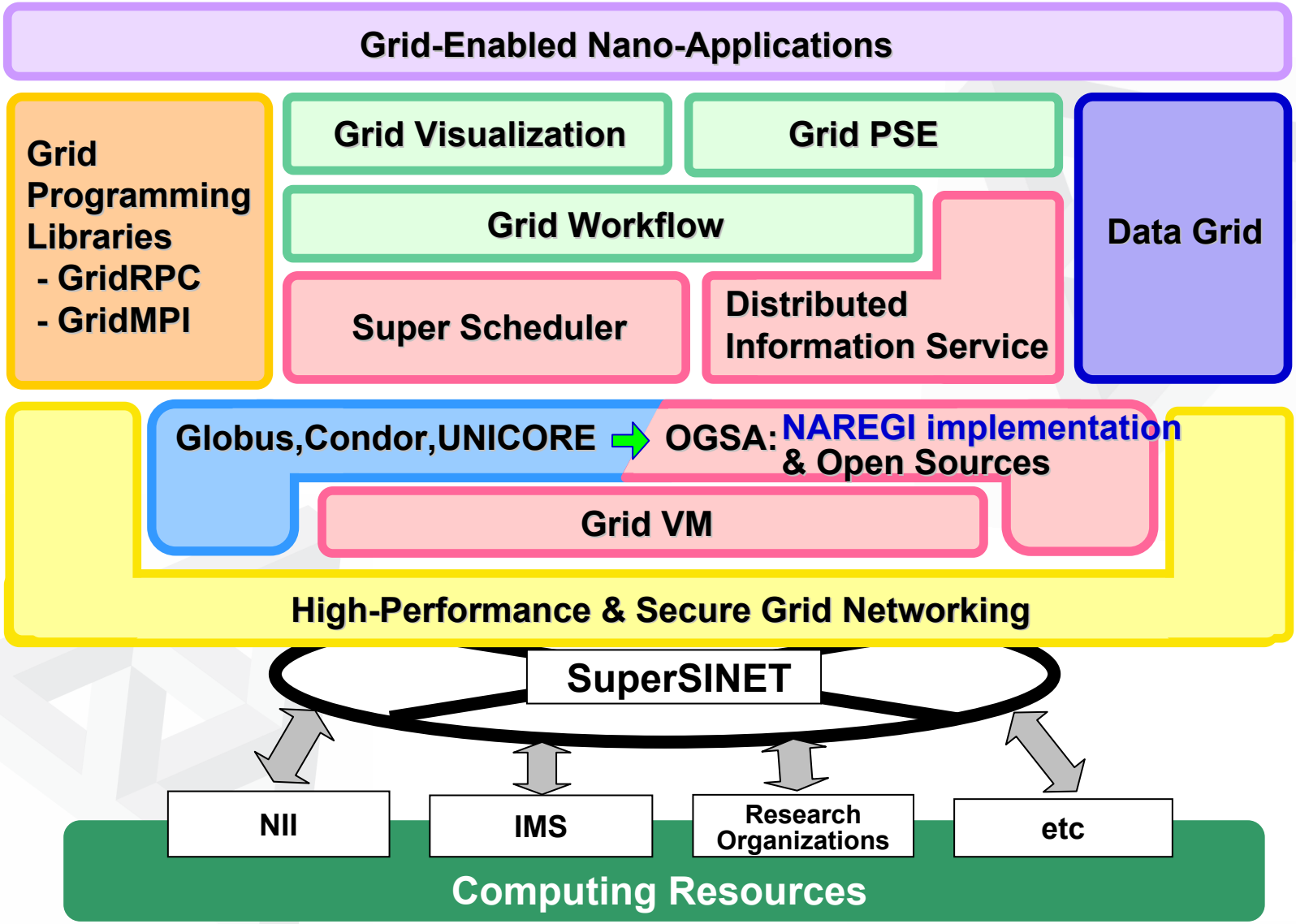
(original slide by Dr. Stephen Pickles, edited by Prof. Matsuoka)



Outline of the NAREGI Project

- NAREGI: National Research Grid Initiative in Japan
- Funded by Japanese Government (MEXT)
- Started from April 2003
- Two main sites:
  - R&D: NII (National Institute for Informatics)
  - Nano Science Application: IMS (Institute for Molecular Science).

NAREGI Software Stack



NAREGI Members

Project Leader: Ken Miura (NII)

(WP1) Resource Management in the Grid Environment
Satoshi Matsuoka (Tokyo Inst. Tech.)

(WP2) Grid Programming Environment
Satochi Sekiguchi, Yoshio Tanaka (AIST)

(WP3) Grid Application Environment
Hitohide Usami (NII), Shigeo Kawata (Utsunomiya Univ.)

(WP4) Data Grid Environment
Hideo Matsuda (Osaka Univ.)

(WP5) High-performance & Security Grid Networking
Shinji Shimojo(Osaka Univ.), Yuji Oie (Kyushi Inst. Tech.),
Makoto Imase (Osaka Univ.)

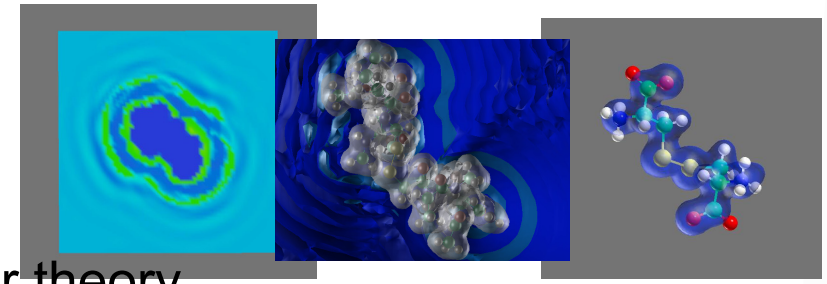
(WP6) Grid-Enable Nano Applications
Mutsumi Aoyagi (Kyushu Univ.)

Nano-Science : coupled simulations on the Grid as the sole future for true scalability

… between Continuum & Quanta.

Material physics
(Infinite system)
- Fluid dynamics
- Statistical physics
- Condensed matter theory

…



Molecular Science
- Quantum chemistry
- Molecular Orbital method
- Molecular Dynamics

…

10^{-6} m 10^{-9}

Limit of Idealization

Multi-Physics

Limit of Computing Capability

Old HPC environment:
- decoupled resources,
- limited users,
- special software, ...

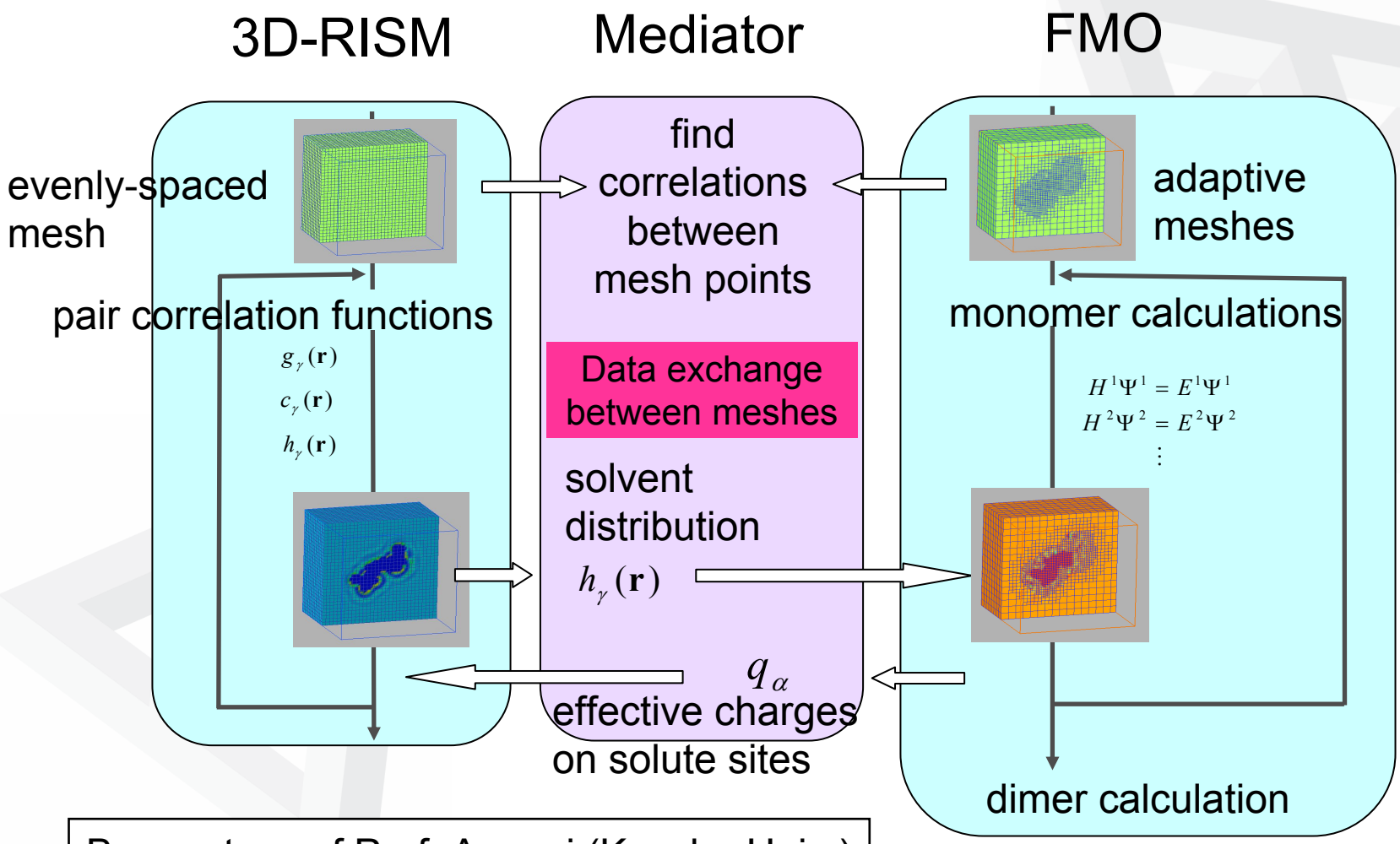
Coordinates decoupled resources;

Meta-computing,
High throughput computing,
Multi-Physics simulation
w/ components and data from different groups
within VO composed in real-time

The only way to achieve true scalability!

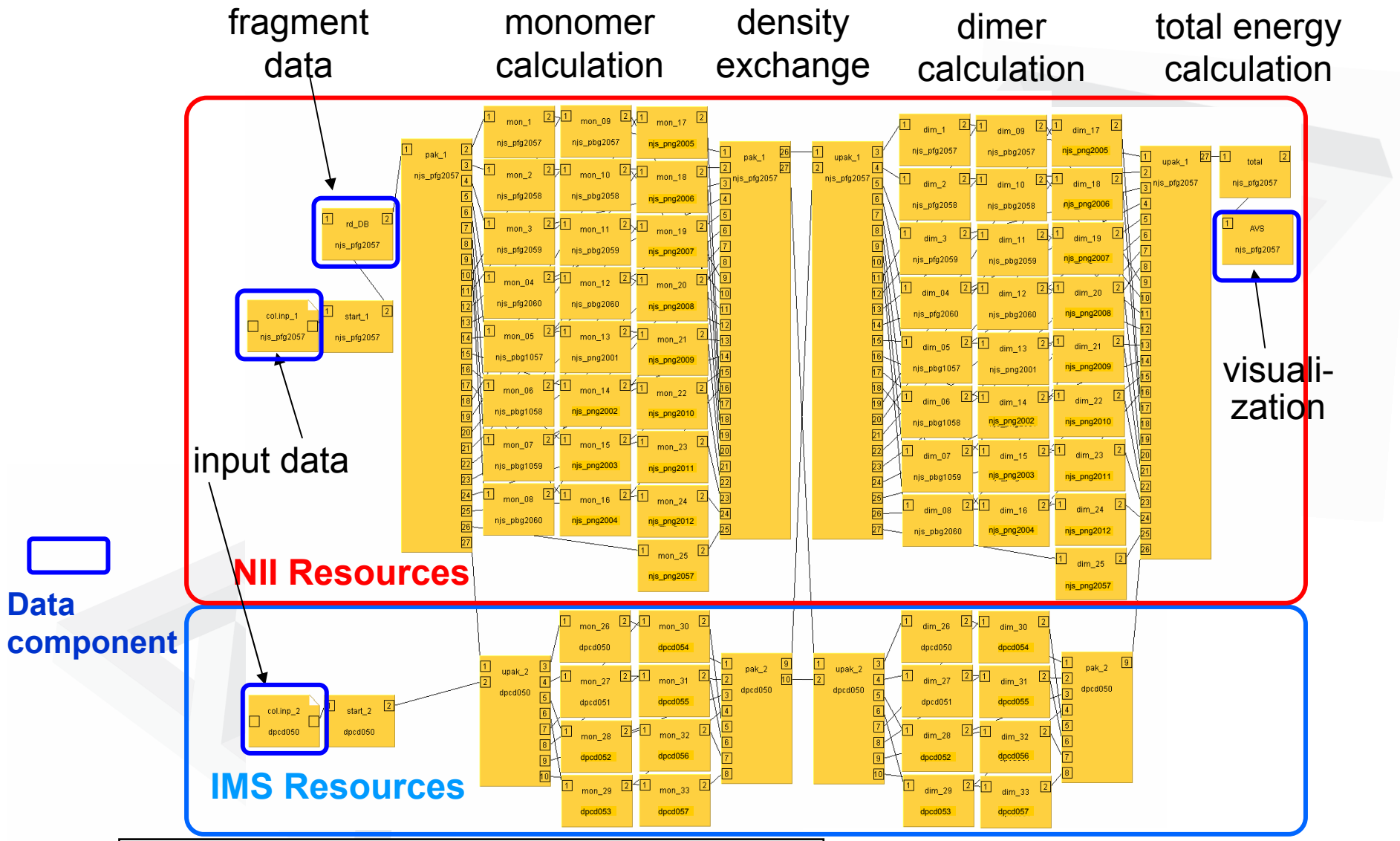
NAREGI Application: Nanoscience

Simulation Scheme



By courtesy of Prof. Aoyagi (Kyushu Univ.)

Workflow based Grid FMO Simulations of Proteins



By courtesy of Prof. Aoyagi (Kyushu Univ.)

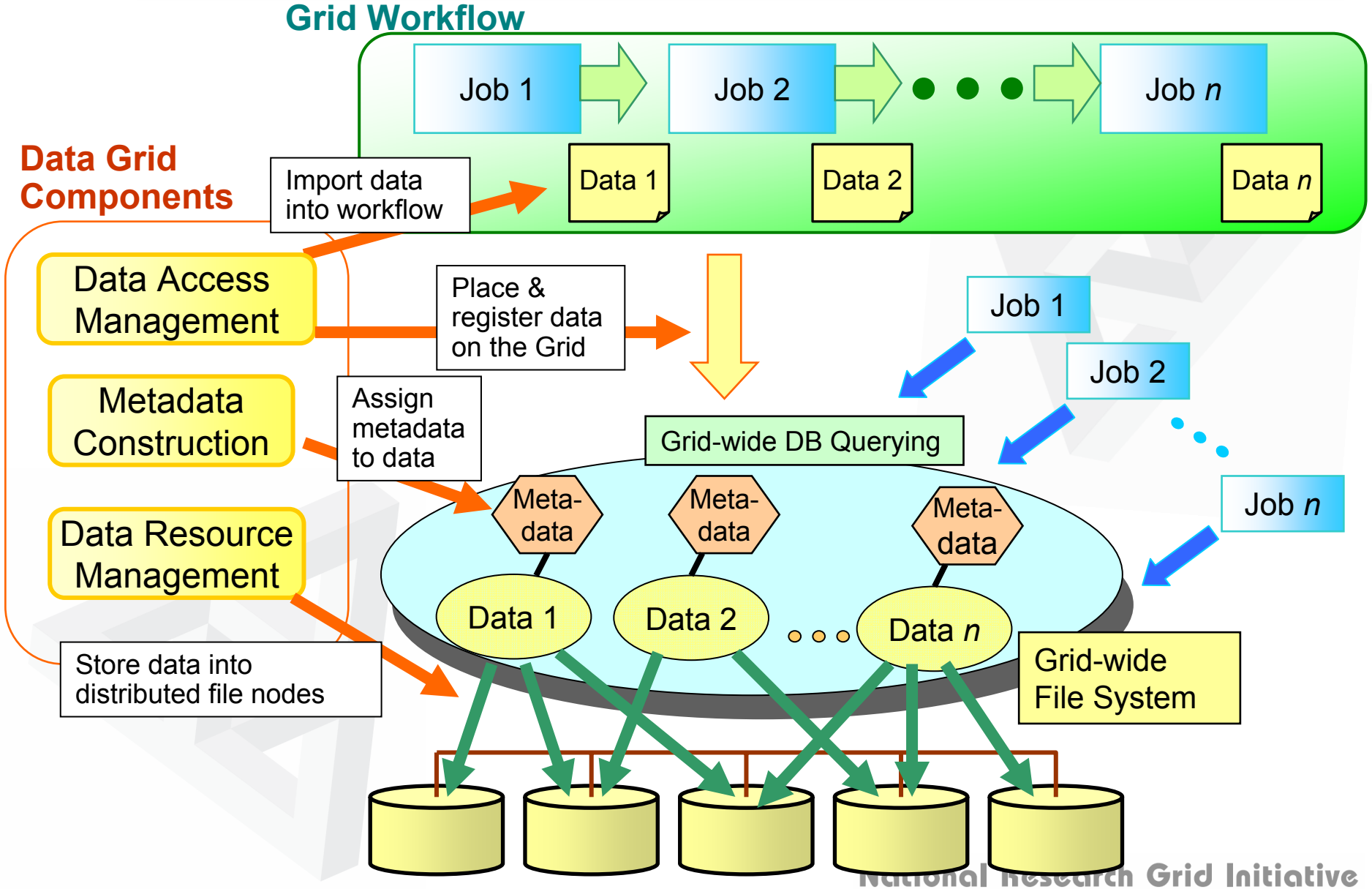
Several Aspects in Data Grid

- Data Transfer
  - GridFTP, RFT, TeraGrid Copy, …
- Data Management
  - SRM, …
- Grid Filesystem
  - SRB, Gfarm, …
- Data Integration
  - OGSA-DAI, myGrid, BRIDGES, …
- Data-intensive Workflow
  - many….

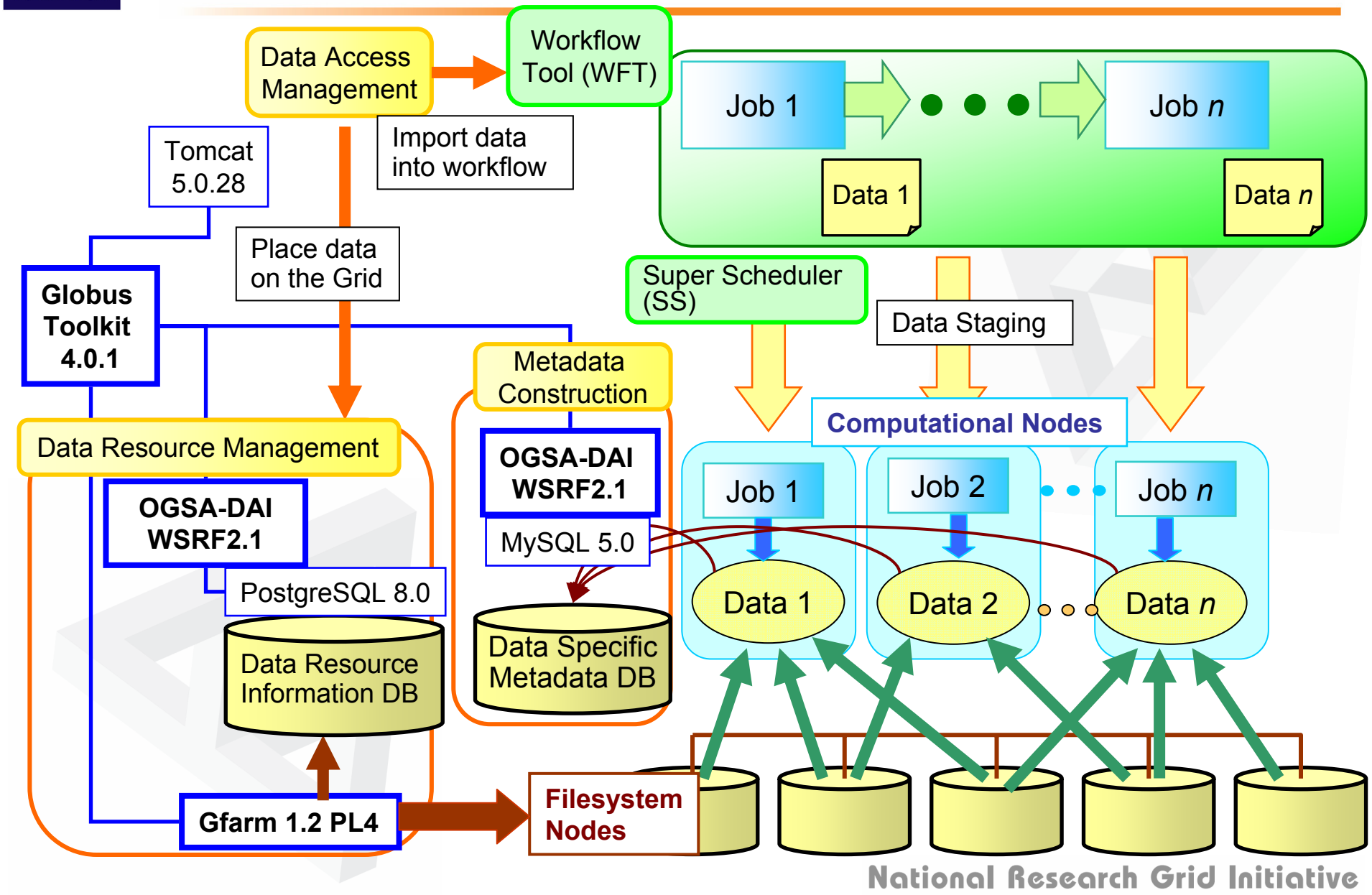
NAREGI DataGrid

- GGF-GFS (AIST Gfarm)
- Data Transfer Service between storage and computation nodes (GridFTP)
- OGSA-DAI for database queries

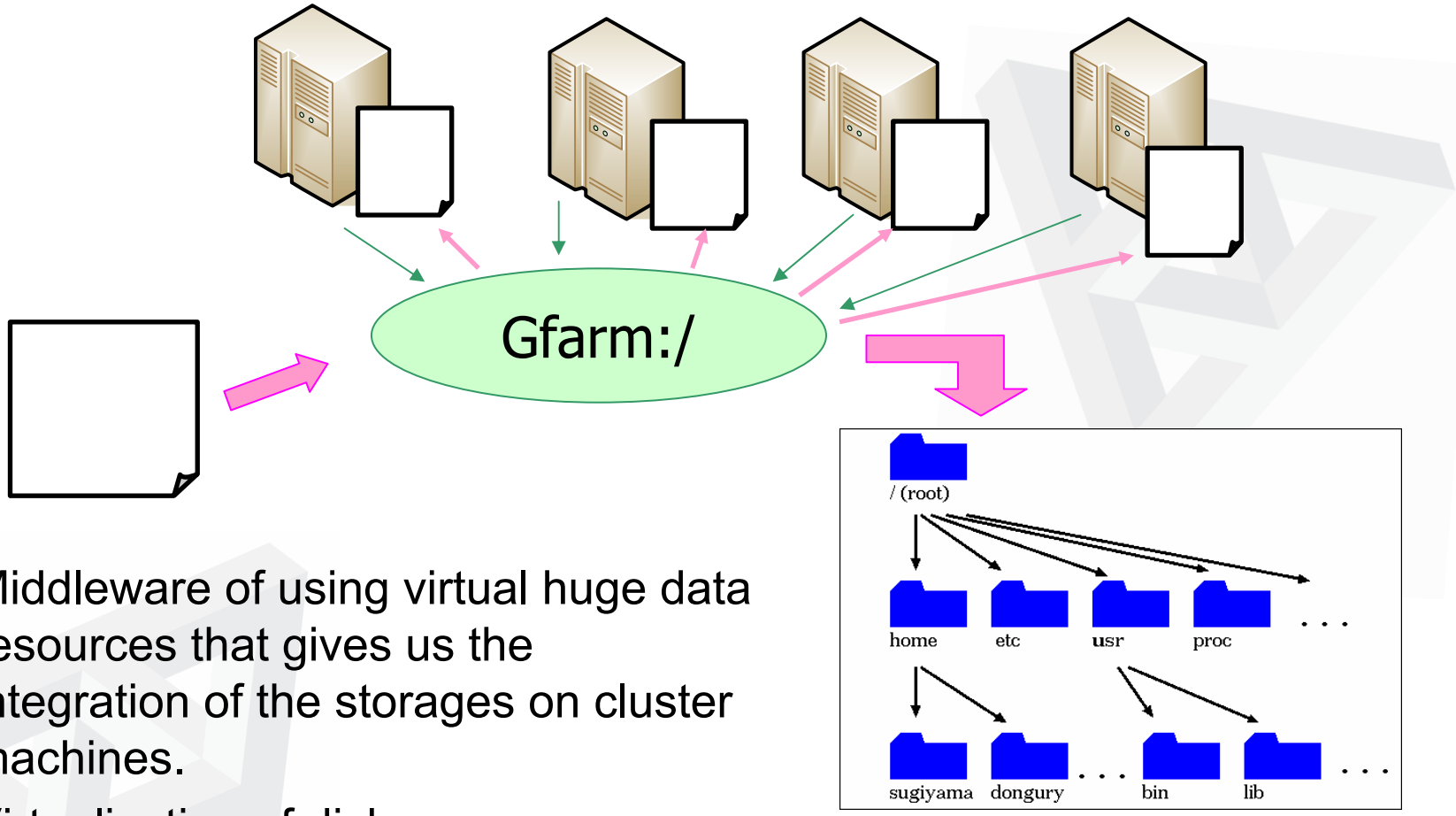
NAREGI Data Grid Environment



Implementation of the Data Grid Environment



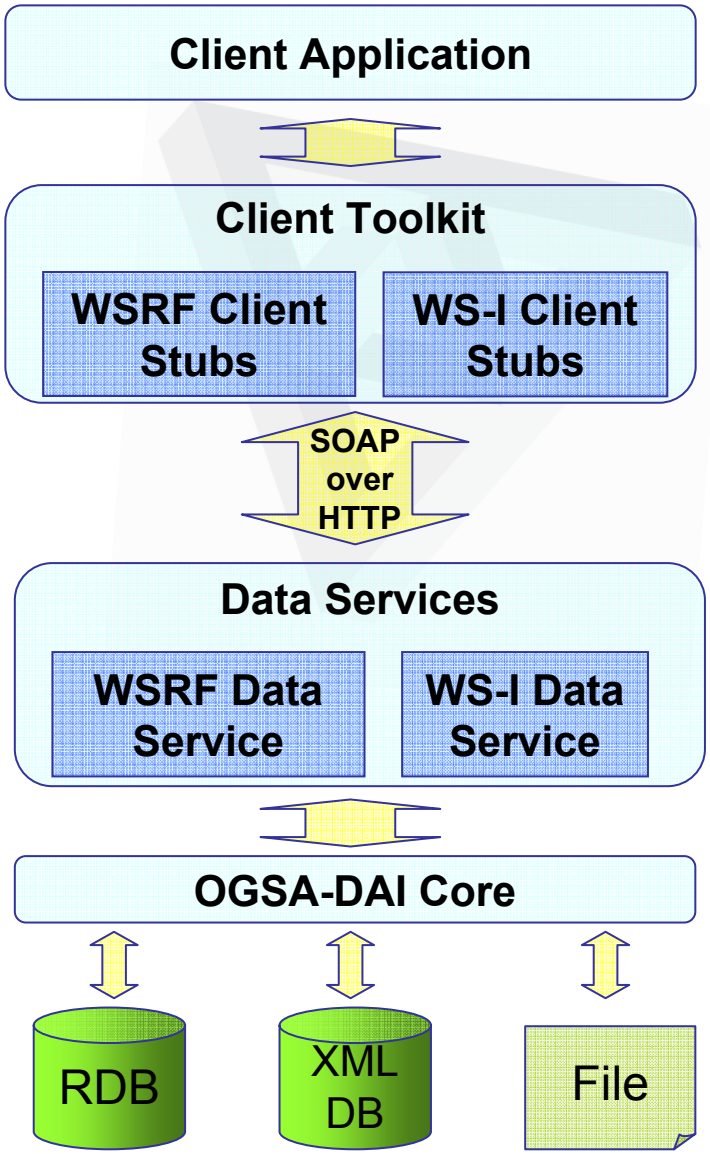
AIST Gfarm: Grid File System



- Middleware of using virtual huge data resources that gives us the integration of the storages on cluster machines.
- Virtualization of disk resources
- Split allocation of huge file

OGSA-DAI

- OGSA-DAI（Data Access and Integration）
- Middleware to facilitate access to data resources
- Data resources are sources/sinks of data
  - OGSA-DAI currently supports:
    - Relational databases
    - XML data resources
    - Files
- Provides a partial virtualisation of data
  - Hide connection mechanism
  - Move computation to the data
  - Do not hide the underlying data model (relational, XML, file)



Scientific Data and Grid

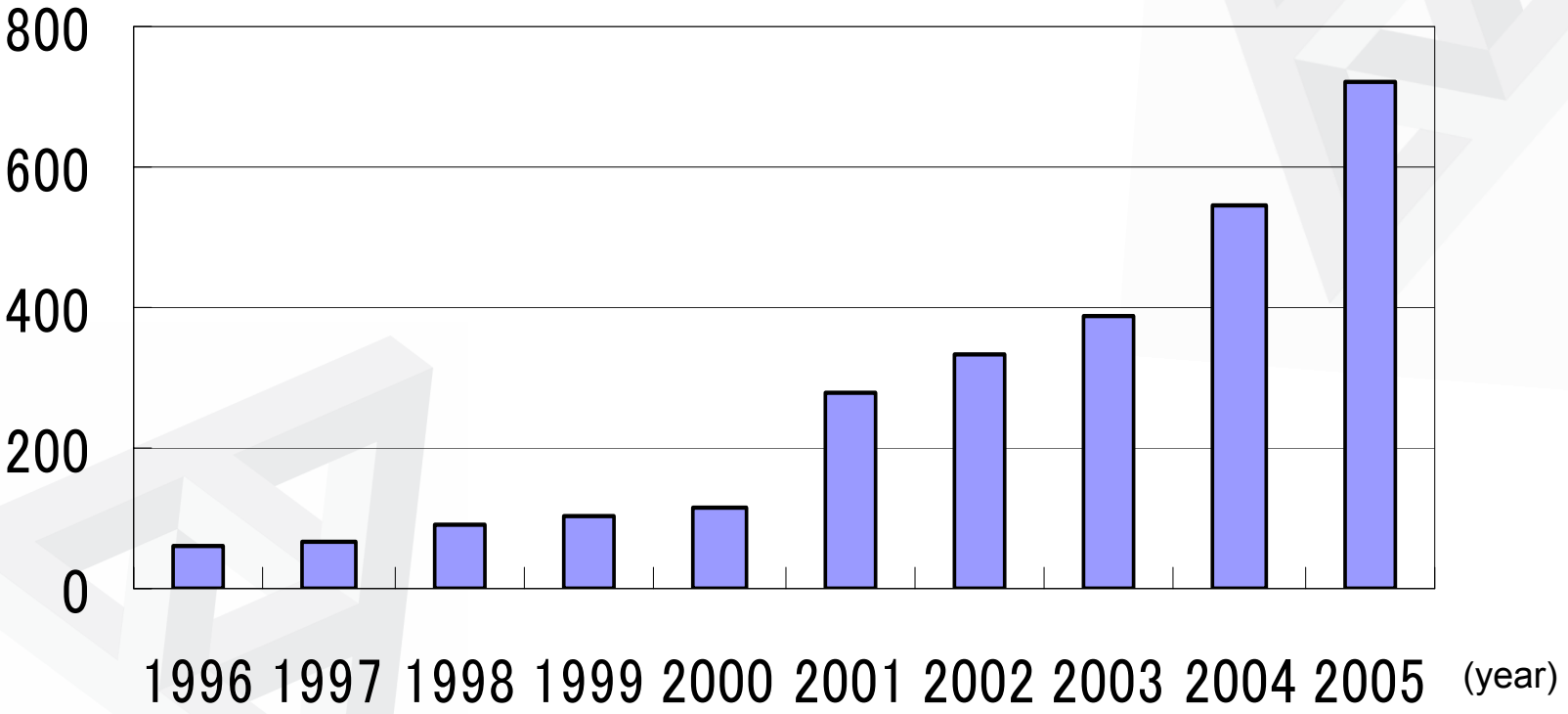
- High Energy Physics (LHC, D-Grid, HEPGrid, ….)
  - Very huge amount of data
  - High speed & reliable data transfer
- Astronomy (Sloan Digital Sky Survey, AstroGrid, …)
  - Virtual Observatory: Large amount of image data
  - Heterogeneous observatories and instruments
  - Distributed query processing

Lifescience DBs

- Not so huge amount of data compared to high-energy physics, astronomy, etc.
- Large number of DBs (> 700 DBs)
- Highly heterogeneous and complex in data description
- Need some semantics

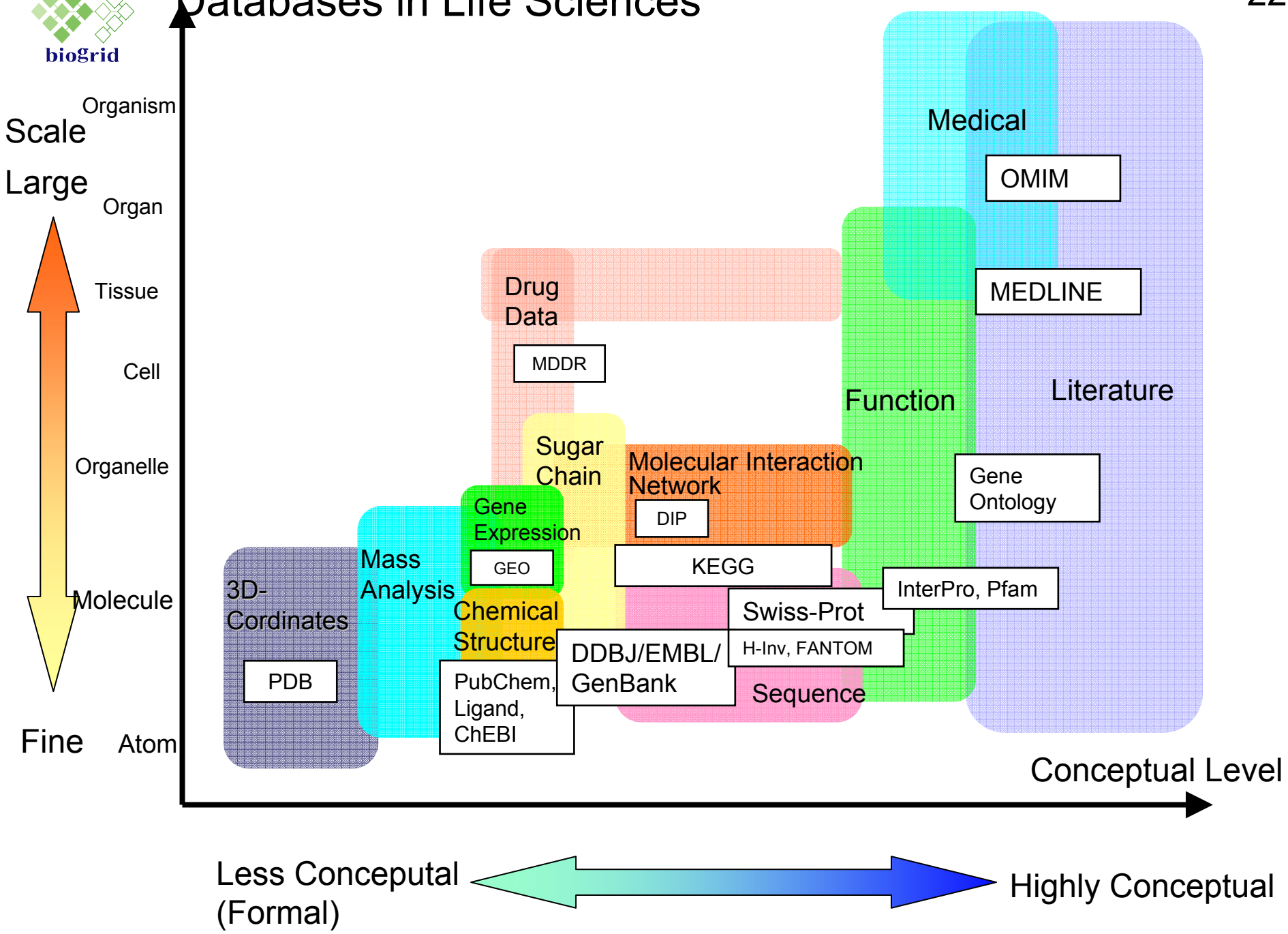
Number of Scientific DBs

- The number of scientific DBs (especially, in life sciences) is very rapidly increasing.



# of life science DBs in Nucleic Acids Research DB issue.

Databases in Life Sciences



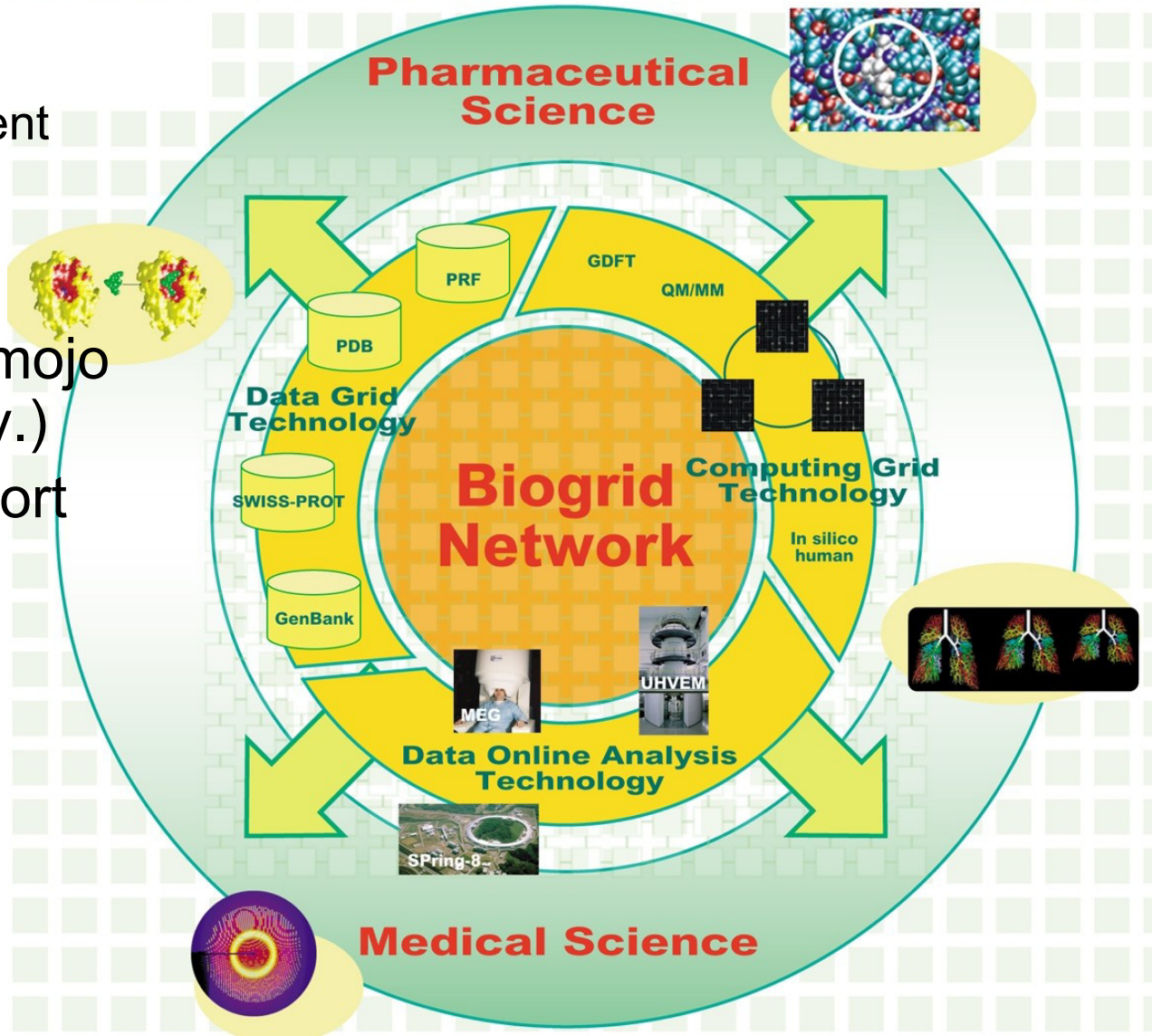
Osaka-U BioGrid Project

- Started from 2002
- Goals

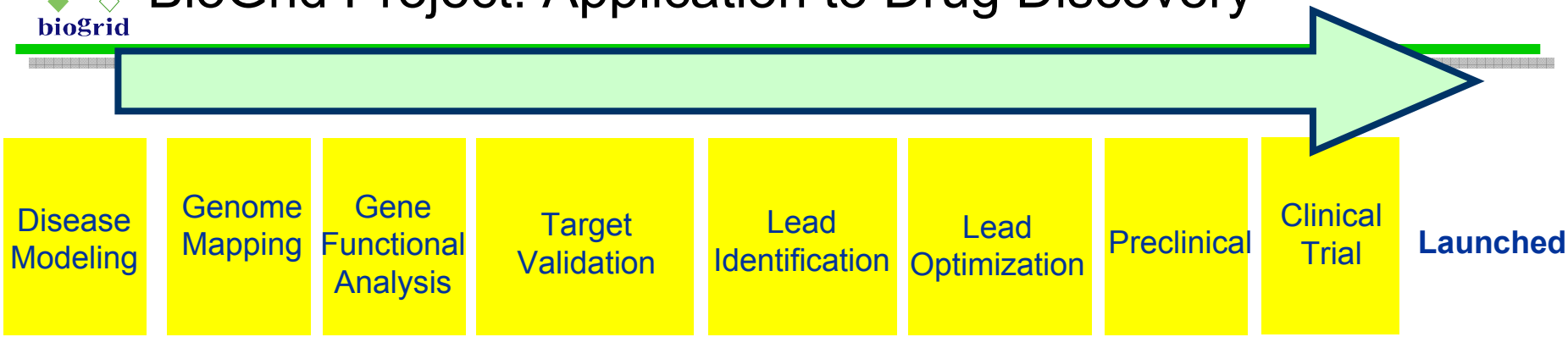
Technology Development for: Pharmaceutical (drug discovery) and Medical Sciences.

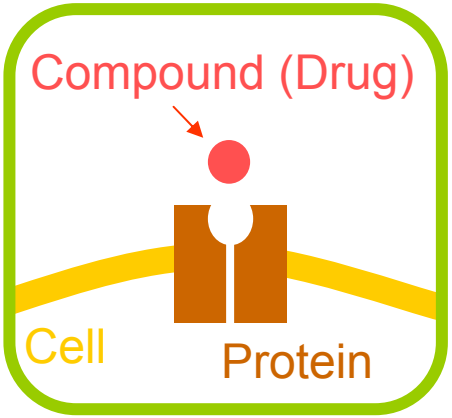
- Leader: Shinji Shimojo (CMC, Osaka Univ.)
- Government Support (MEXT): 5years (1~4M$/year)
- Web site:

http://www.biogrid.jp



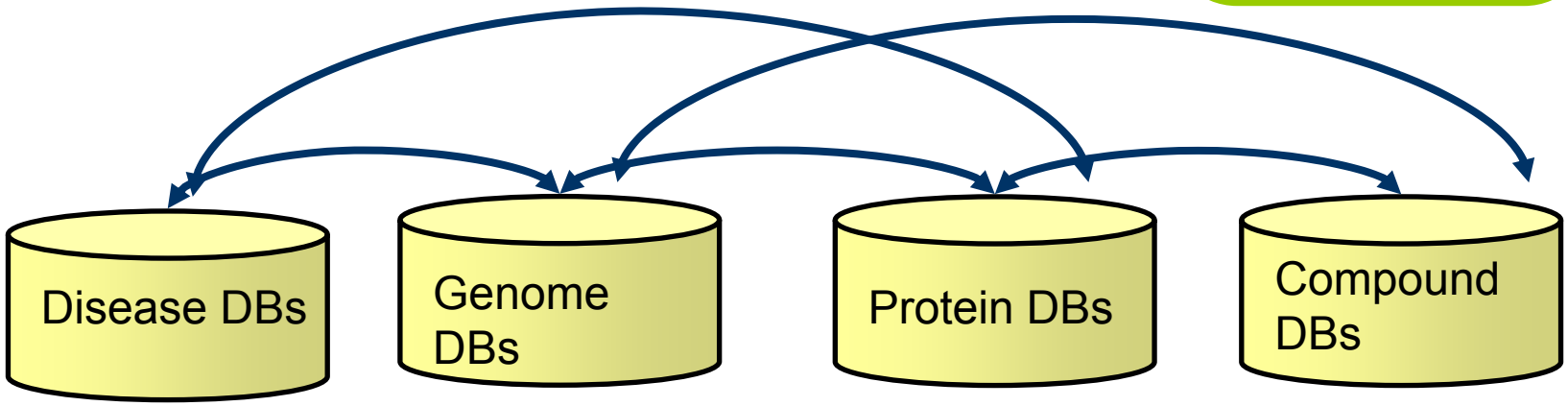
BioGrid Project: Application to Drug Discovery



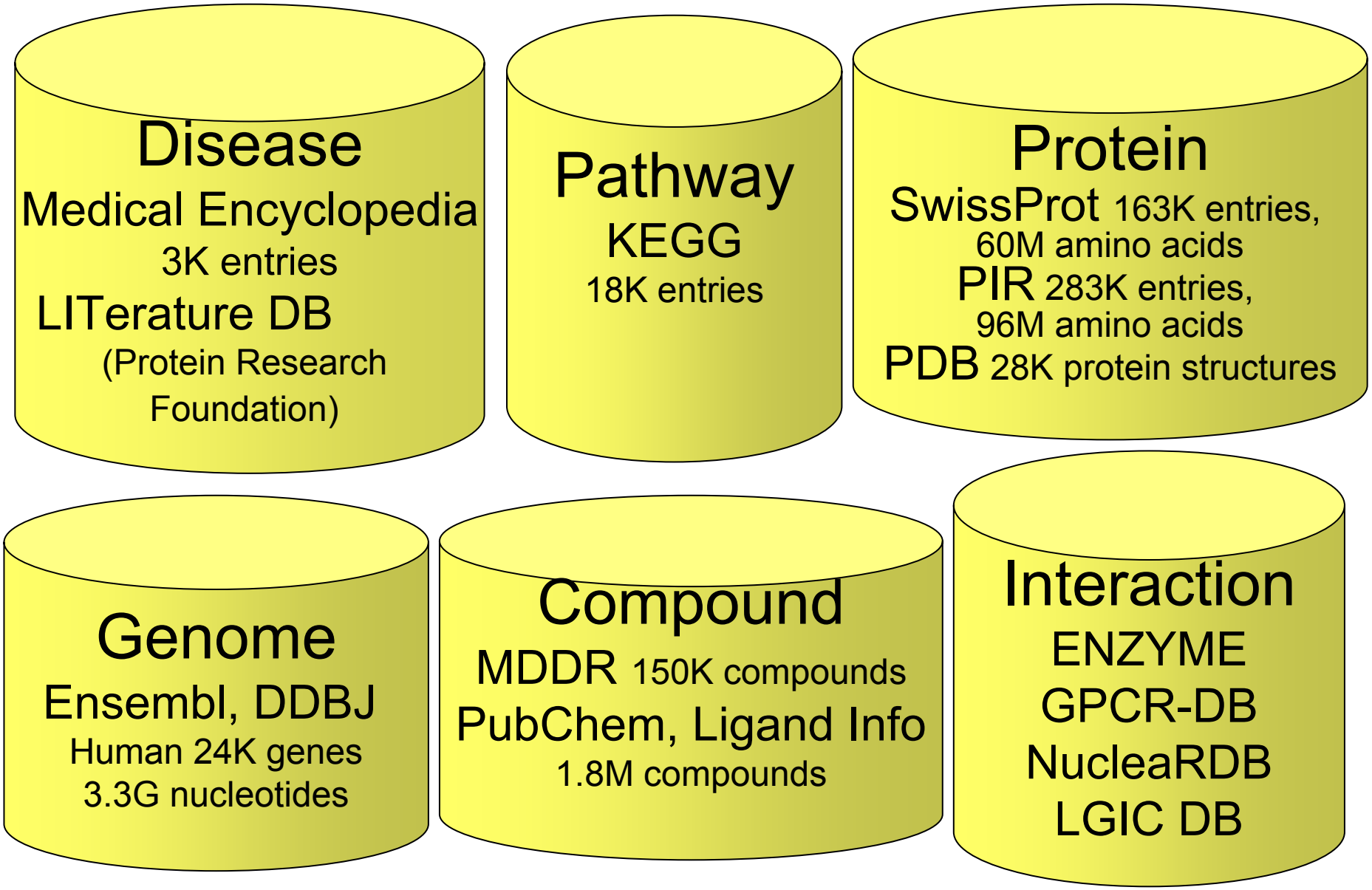


Data Grid Issues:

1. Heterogeneity of data descriptions.
2. Large number of relationships



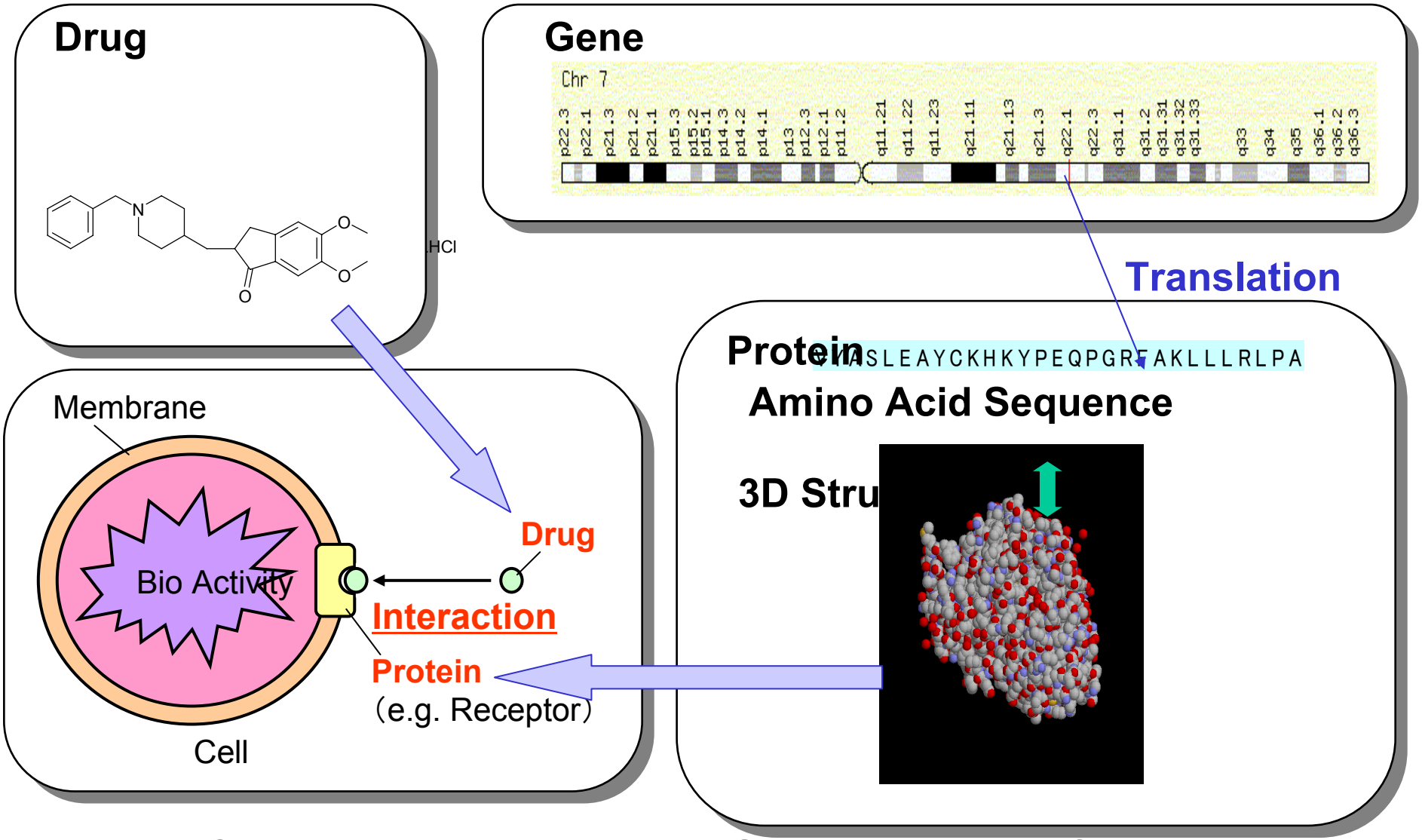
Databases in BioGrid Project



Increase of Public Chemical Databases

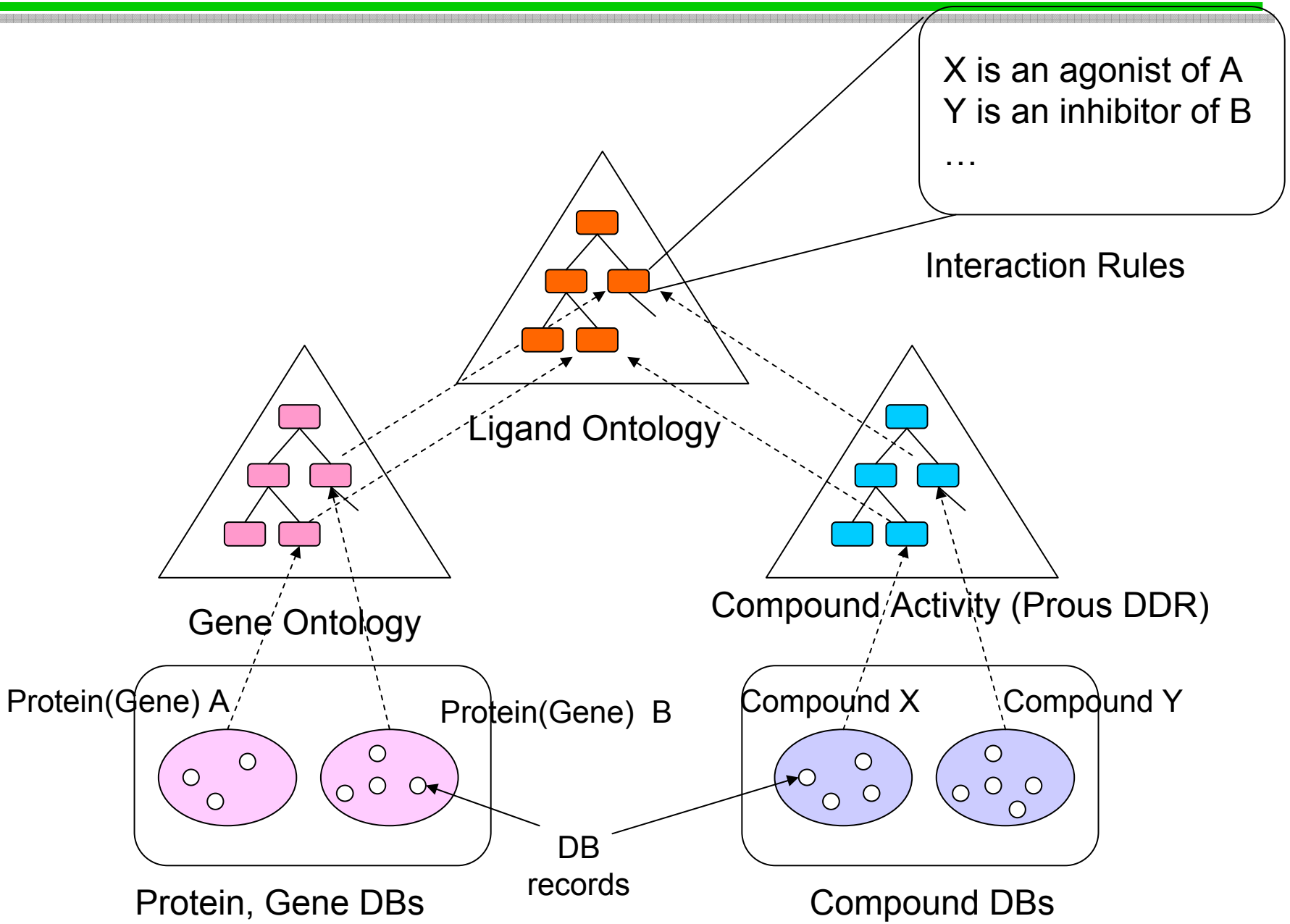
- Ligand.Info（http://ligand.info/）：composed of the following data.
  - ChemBank subset 2,344 records
  - ChemPDB subset 4,009 records
  - KEGG subset 10,005 records
  - Anti-HIV NCE subset 42,689 records
  - Drug-likeness NCI subset 192,323 records
  - Not annotated NCI subset 15,237 records
  - AKos GmbH subset 544,391 records
  - Ligand.InfoAsinex Ltd. subset 348,276 records
  - Tim Tec subset 7,500 records (in total: 1,159,274 records)
- PubChem（http://pubchem.ncbi.nlm.nih.gov/）
  - 711,361 records
- Drug Bank(http://redpoll.pharmacy.ualberta.ca/drugbank)
  - 256 records (company names)

Protein-Compound Interaction

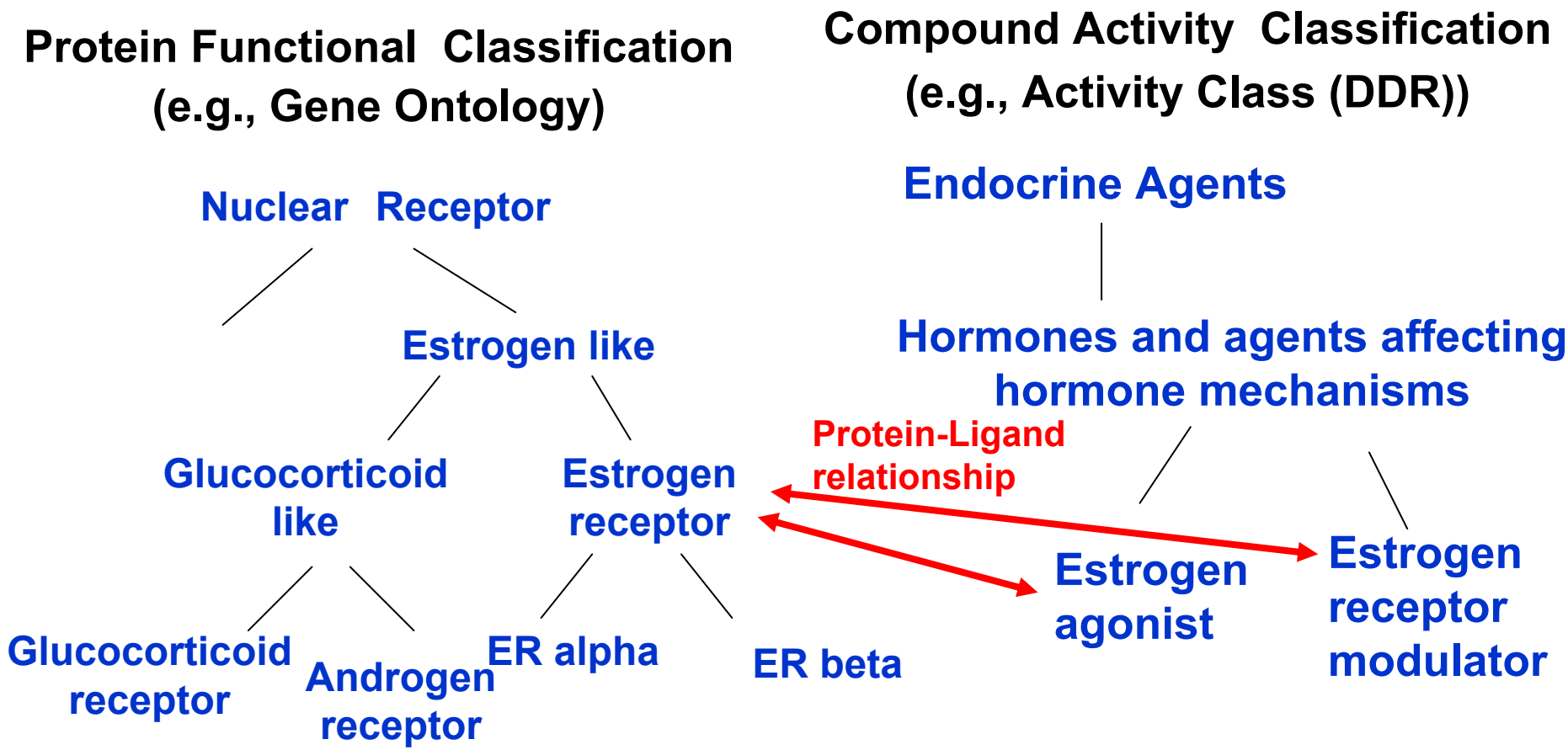


Protein-Compound Interaction Search is one of the most important technologies in drug discovery.

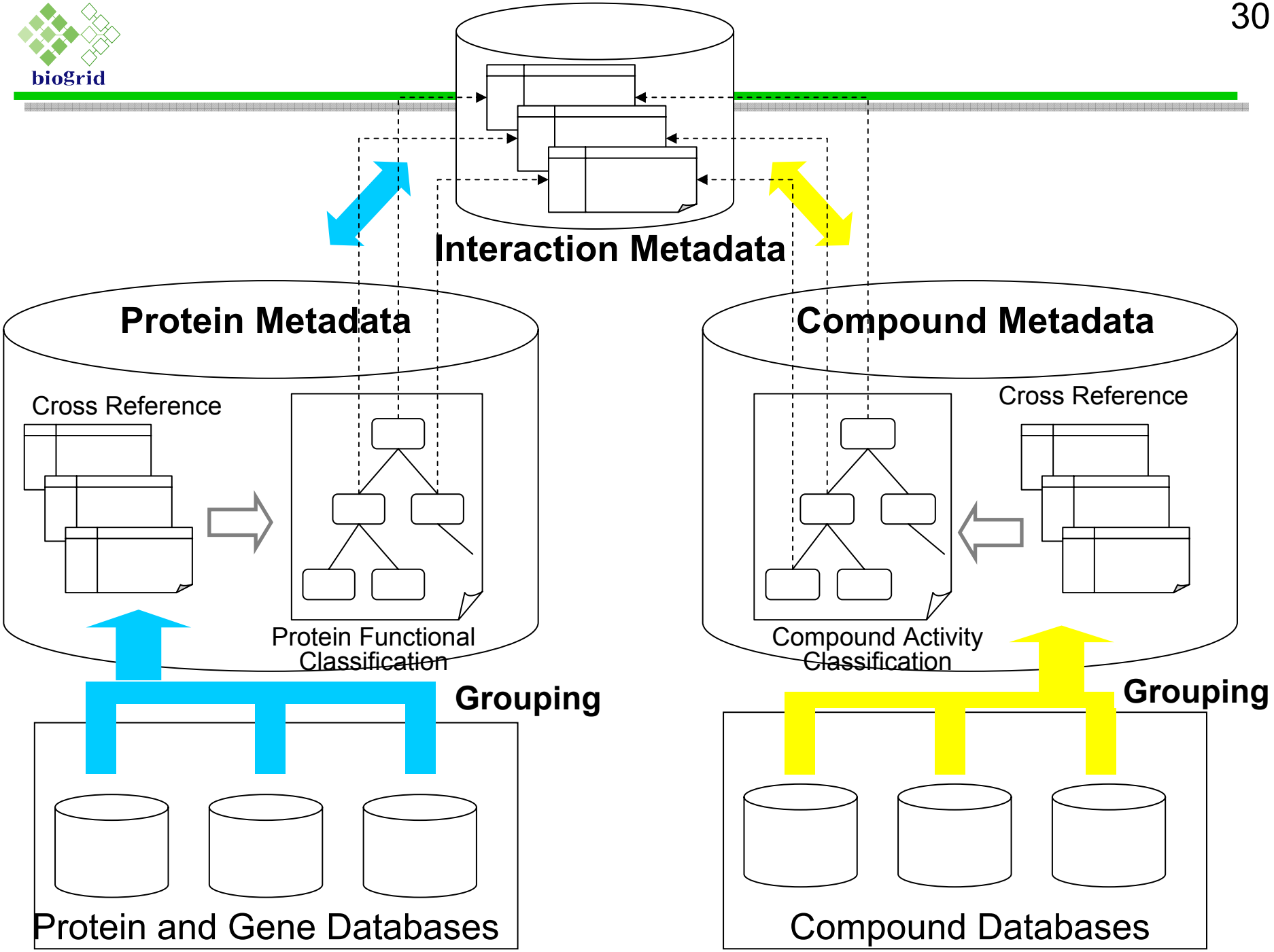
Ontology-based Metadata



Relationship between Protein Function and Drug Activity

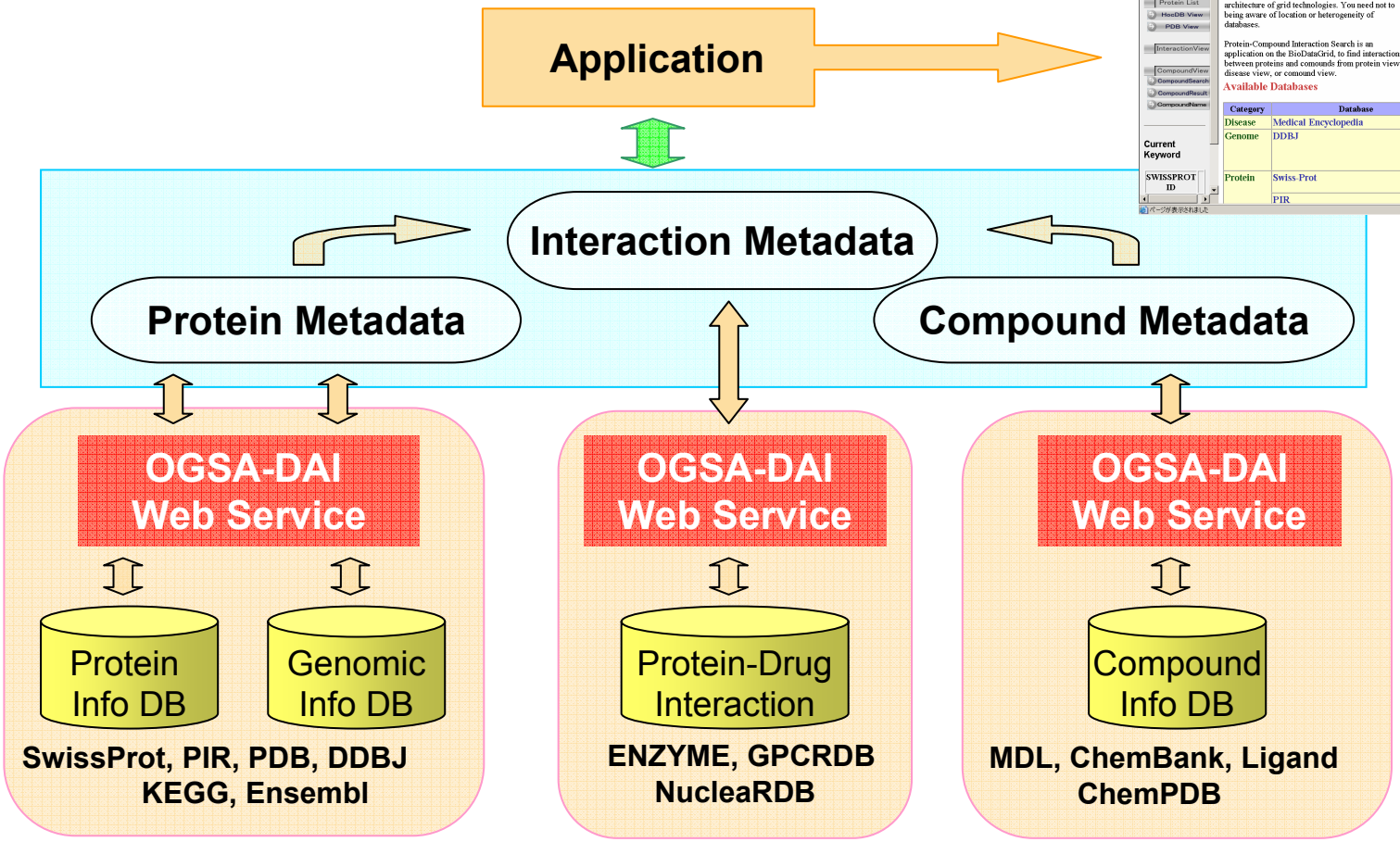


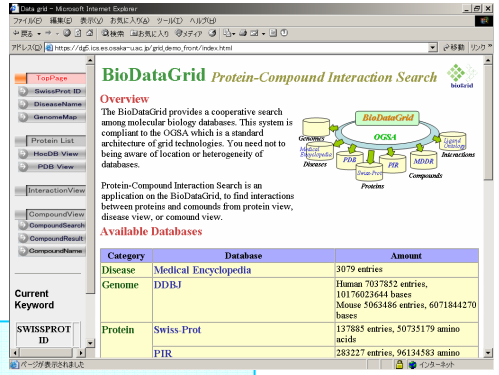
- Extract relationships between protein functions and compound activities by traversing hierarchical classifications (or *ontologies*)


biogrid
Interaction Metadata
Protein Metadata
Cross Reference
Protein Functional Classification
Compound Metadata
Cross Reference
Compound Activity Classification
Grouping
Grouping
Protein and Gene Databases
Compound Databases

Outline of Previous Data Integration System

Database federation using Grid technology.





View

Databases for each category are provided as Grid Services

Retrieve Known Compounds Bound to Similar Proteins

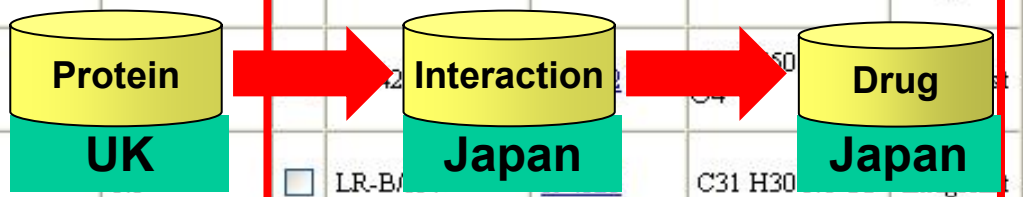
Protein Compound Interaction View

Save Compound | Compound Similarity | Top [ ] Compounds

E-value [ ]

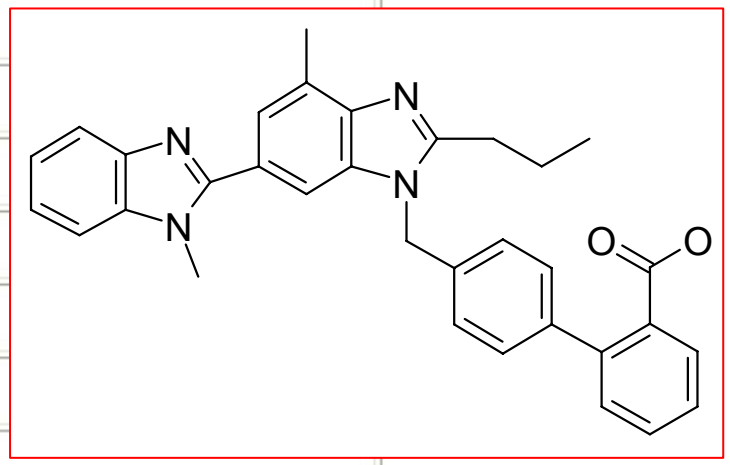
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

SWISS-PROT ID	SWISS-PROT Acc.	PIR	Protein Name	Homology Score	Homology Evalue		Compound	MDDR EXTREG	Molformula	Type
AG2R_HUMAN	P30556	JC1104	Type-1 angiotensin II receptor (AT1) (AT1AR)	657	0.0	☐	212740	212740	C26 H25 N7 O	antagonist
AG2R_HUMAN	P30556	JC1104	Type-1 angiotensin II receptor (AT1) (AT1AR)	657	0.0	☐	260825	260825	C28 H30 N8 O5	antagonist
AG2R_HUMAN	P30556	JC1104	Type-1 angiotensin II receptor (AT1) (AT1AR)	657	0.0	☐	317215	317215	C27 H24 F3 N5 O3 S	antagonist
AG2R_HUMAN	P30556	JC1104	Type-1 angiotensin II receptor (AT1) (AT1AR)	657	0.0	☐	FK-739	193909	C24 H22 N7 . Na	antagonist
AG2R_HUMAN	P30556	JC1104	Type-1 angiotensin II receptor (AT1) (AT1AR)	657	0.0	☐	XR-510	211727	C39 H47 F N5 O6 S . K	antagonist
AG2R_HUMAN	P30556	JC1104	Type-1 angiotensin II receptor (AT1) (AT1AR)	657	0.0	☐	UR-7198	225409	C27 H29 N3 O2	antagonist
AG2R_HUMAN	P30556	JC1104	Type-1 angiotensin II receptor (AT1) (AT1AR)	657						
AG2R_HUMAN	P30556	JC1104	Type-1 angiotensin II receptor (AT1) (AT1AR)	657		☐	LR-B/		C31 H30	



Display Information of Known Compound

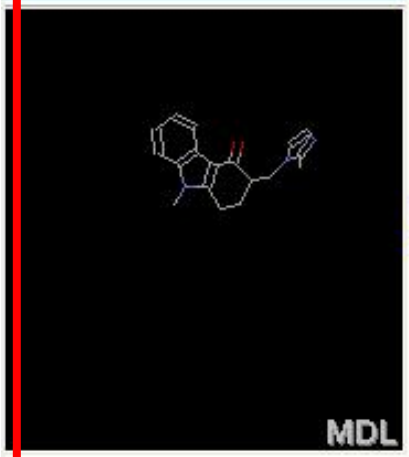
Cas number	144701-48-4
Chemical name	Telmisartan
Company code	
Compound ID	30005957
Molformula	C33H30N4O2
Molname	Telmisartan
Phase	
2D Regno	
Source1	
Source2	
Trademark	Kessar, Noltam, Nolvadex, Nourytam, Tamofen, Tamox
Molecular Weight	514.62
logp	8.486
SMILES	CCCc1nc2c(C)cc(cc2n1Cc3ccc(cc3)-c4ccccc4C(O)=O)-c5nc6ccccc6n5C



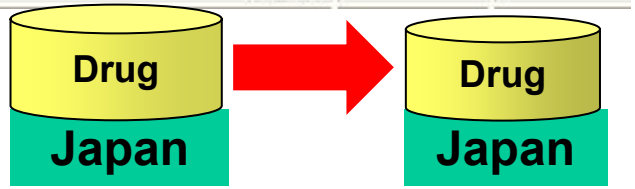
Telmisartan

Drug
Japan

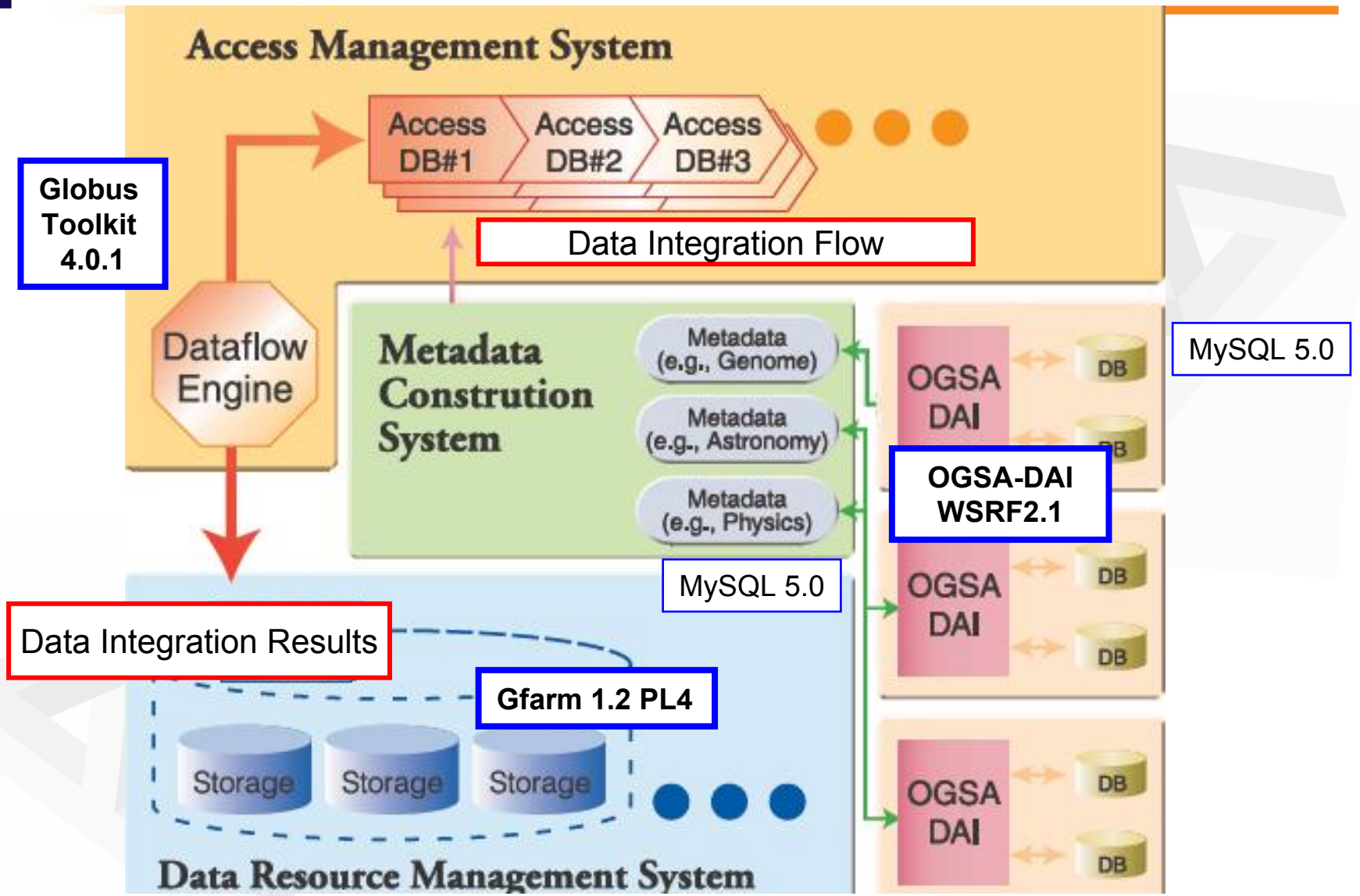
Compound Search Result

Similar Compound				Score	Inquiry Compound		
	Compound Name	Compound ID	Molformula			Compound Name	Compound ID
	Ondansetron hydrochloride	30006044	C18H24ClN3O3	0.686		Telmisartan	30005957
	ro 32-0432	10000554	C28H28N4O2	0.676		Telmisartan	30005957

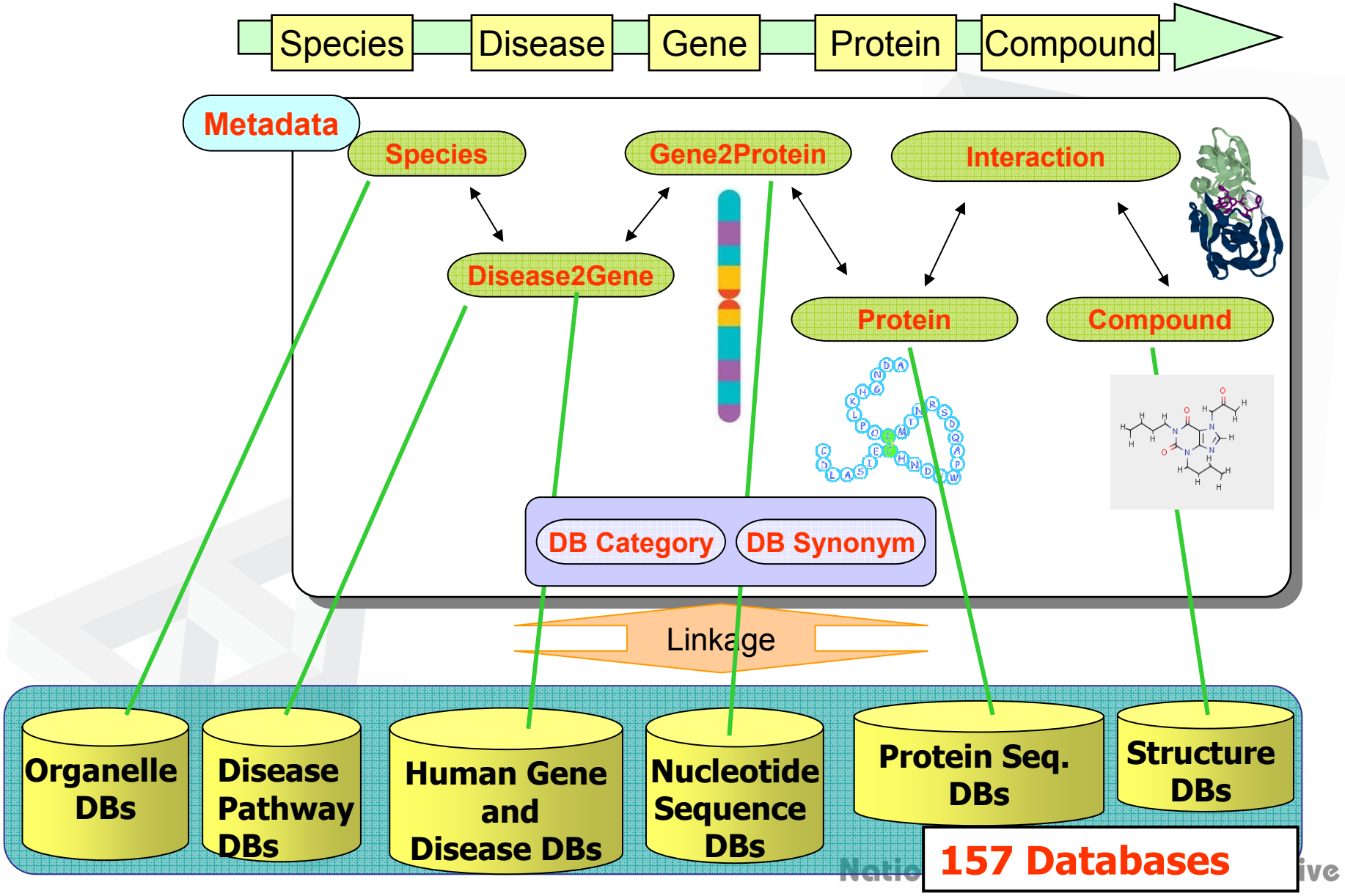
Compounds possibly-interacted to the target protein



NAREGI Data Integration w/ Workflow

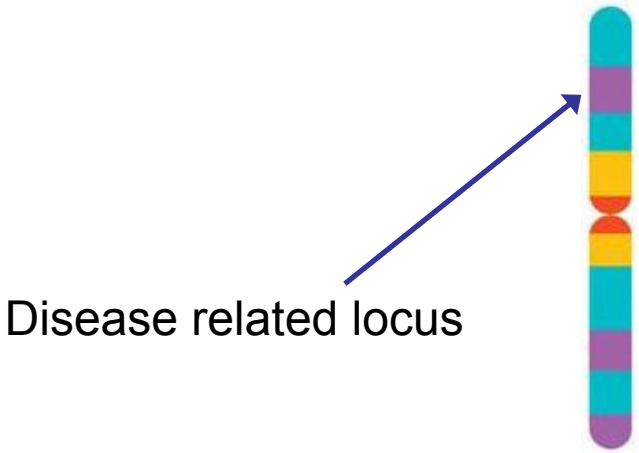


Data Integration Workflow using Metadata



Disease to Gene & Gene to Protein Metadata

- Disease to Gene



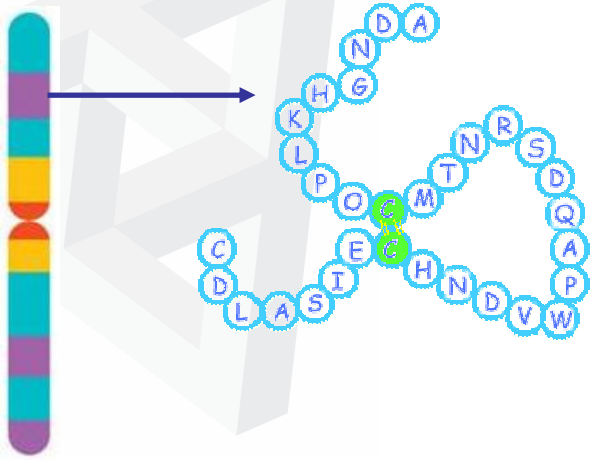
Extraction of the relationships between diseases and genes from OMIM

Example

Hypertension ⇔ PPARG, PPARG1, PPARG2, etc..

Diabete ⇔ IDDM1, IDDM15, IDDM8, IDDM10, etc..

- Gene to Protein



Extraction of translation information from UniGene

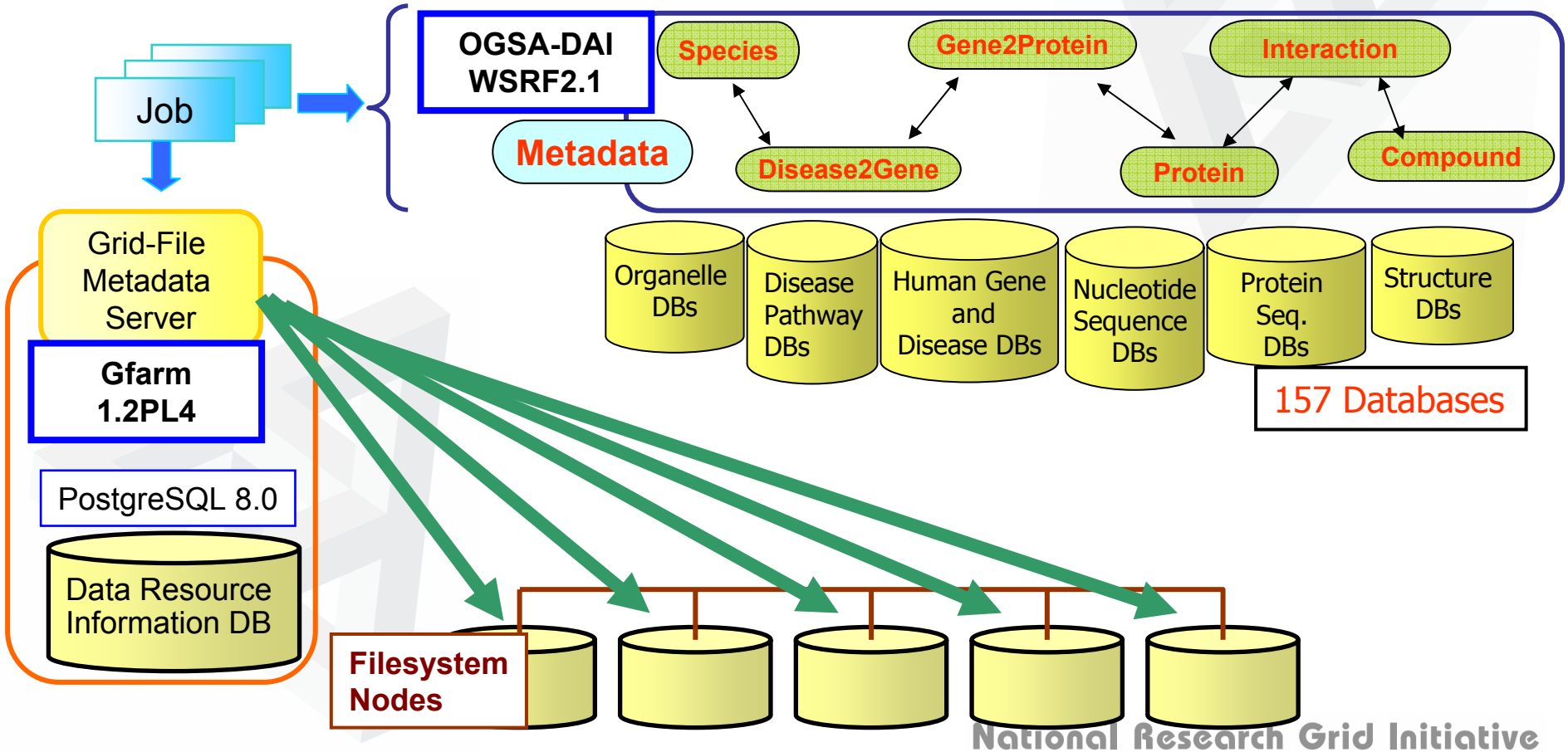
Example.

PPARG ⇔ swiss-prot P41830

PPARG ⇔ swiss-prot P13055

Metadata Management

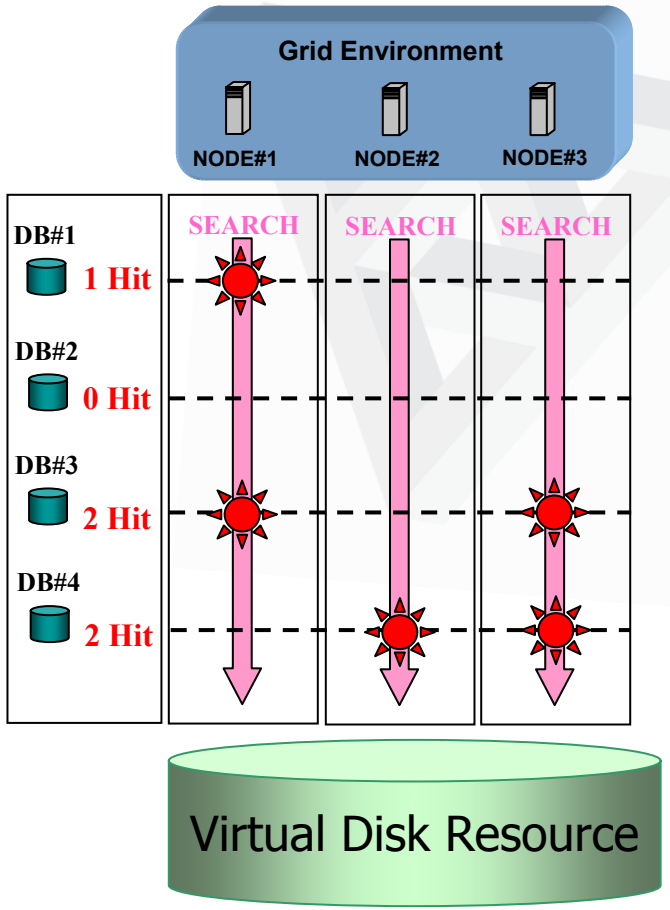
- Two types of metadata
  - Grid file metadata
  - User metadata (domain-specific metadata)



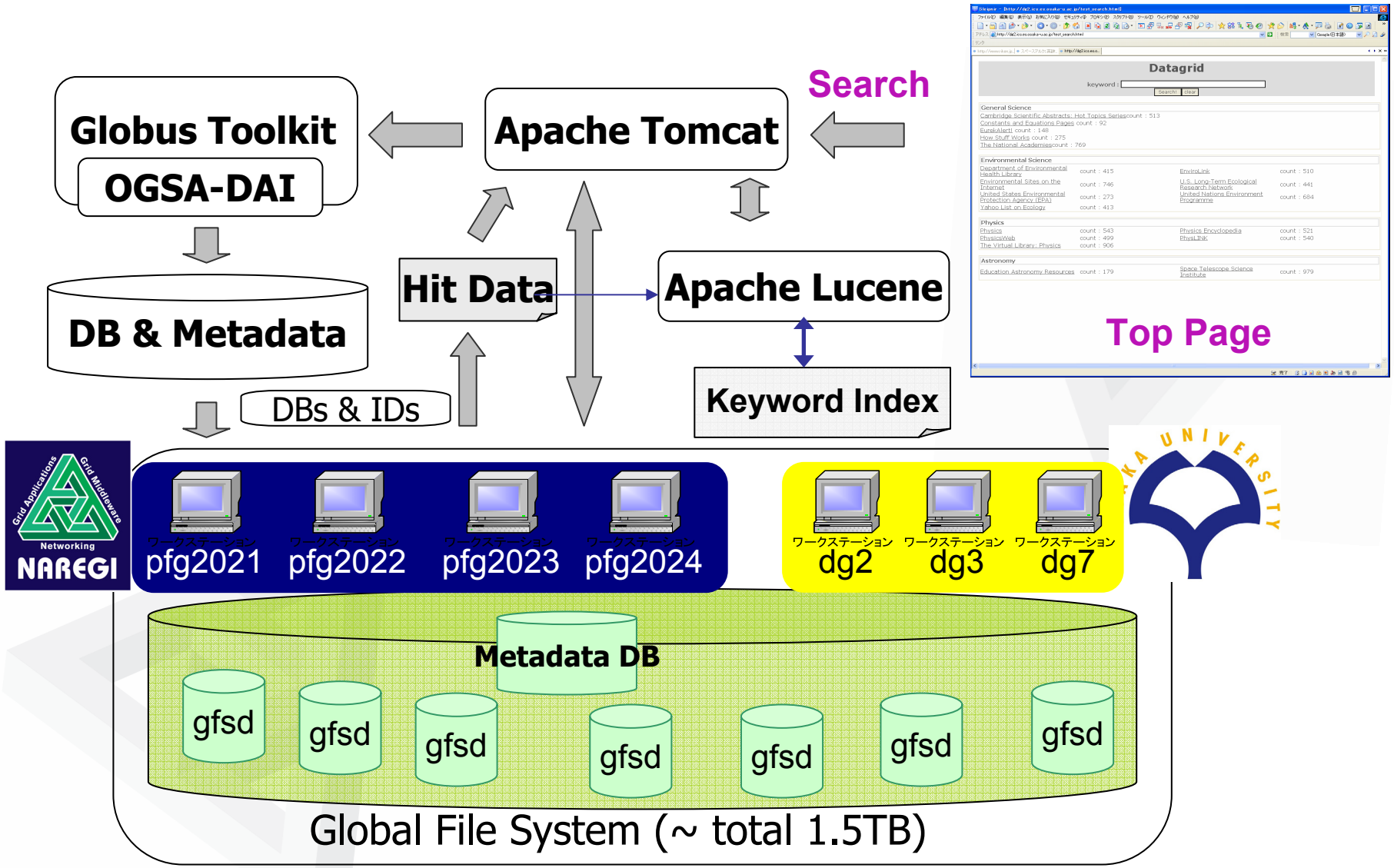
Parallel Search across DBs

- **Retrieve in parallel against lifescience DBs by using Globus Toolkit, OGSA-DAI, and Gfarm.**
- **We extract the information of related records among different lifescience DBs and integrate their results.**

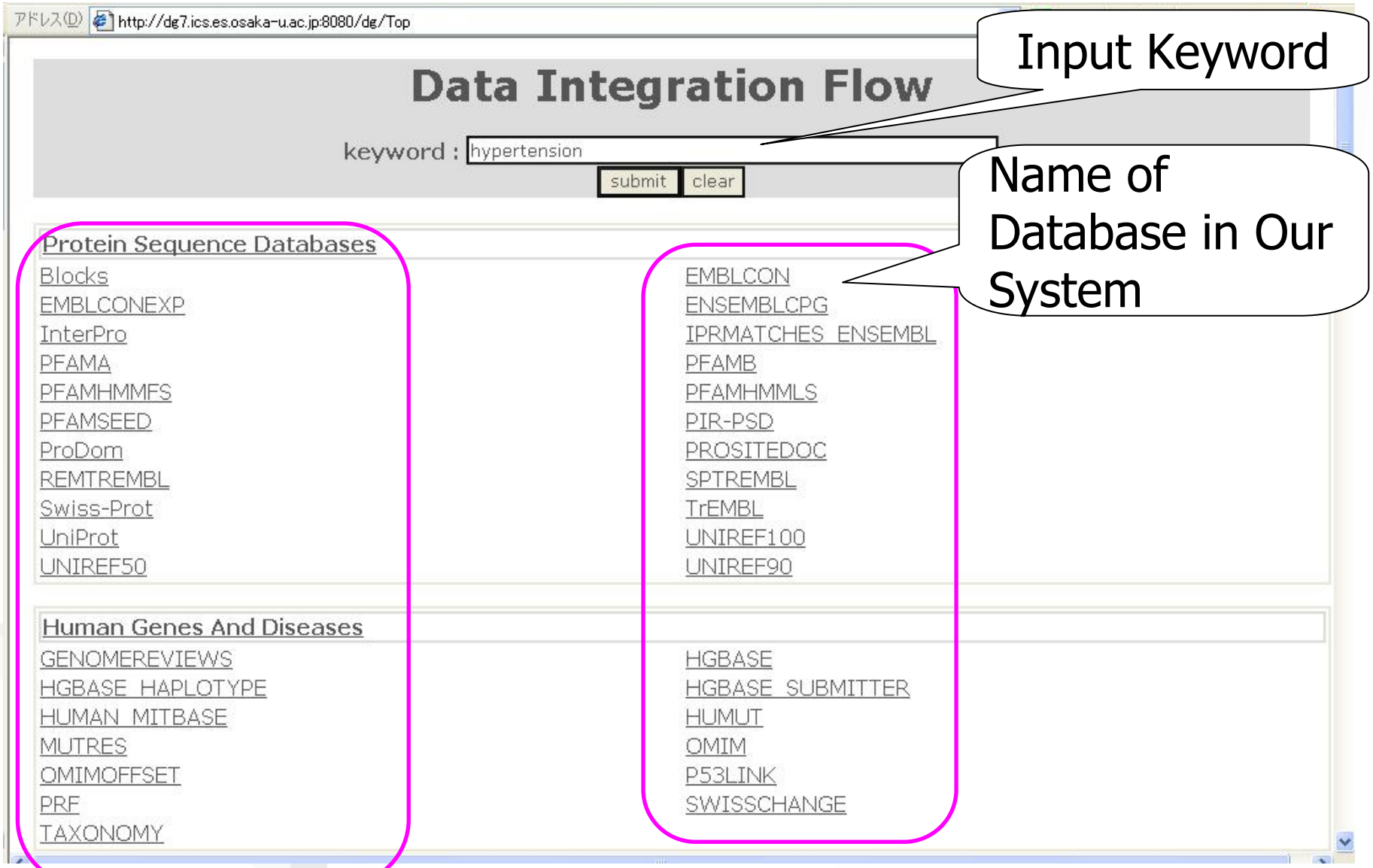
▼Search across Databases



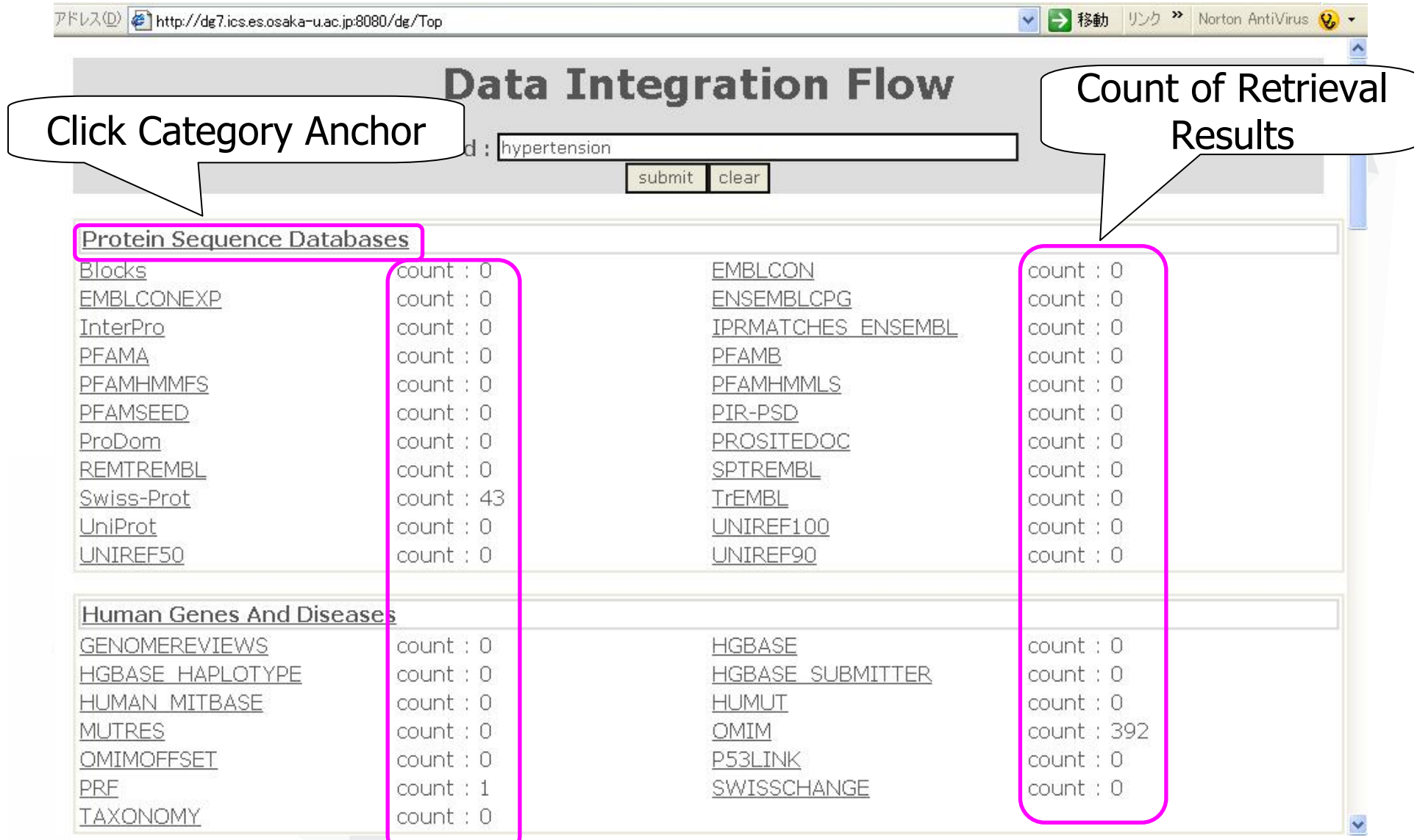
Data Integration System Overview



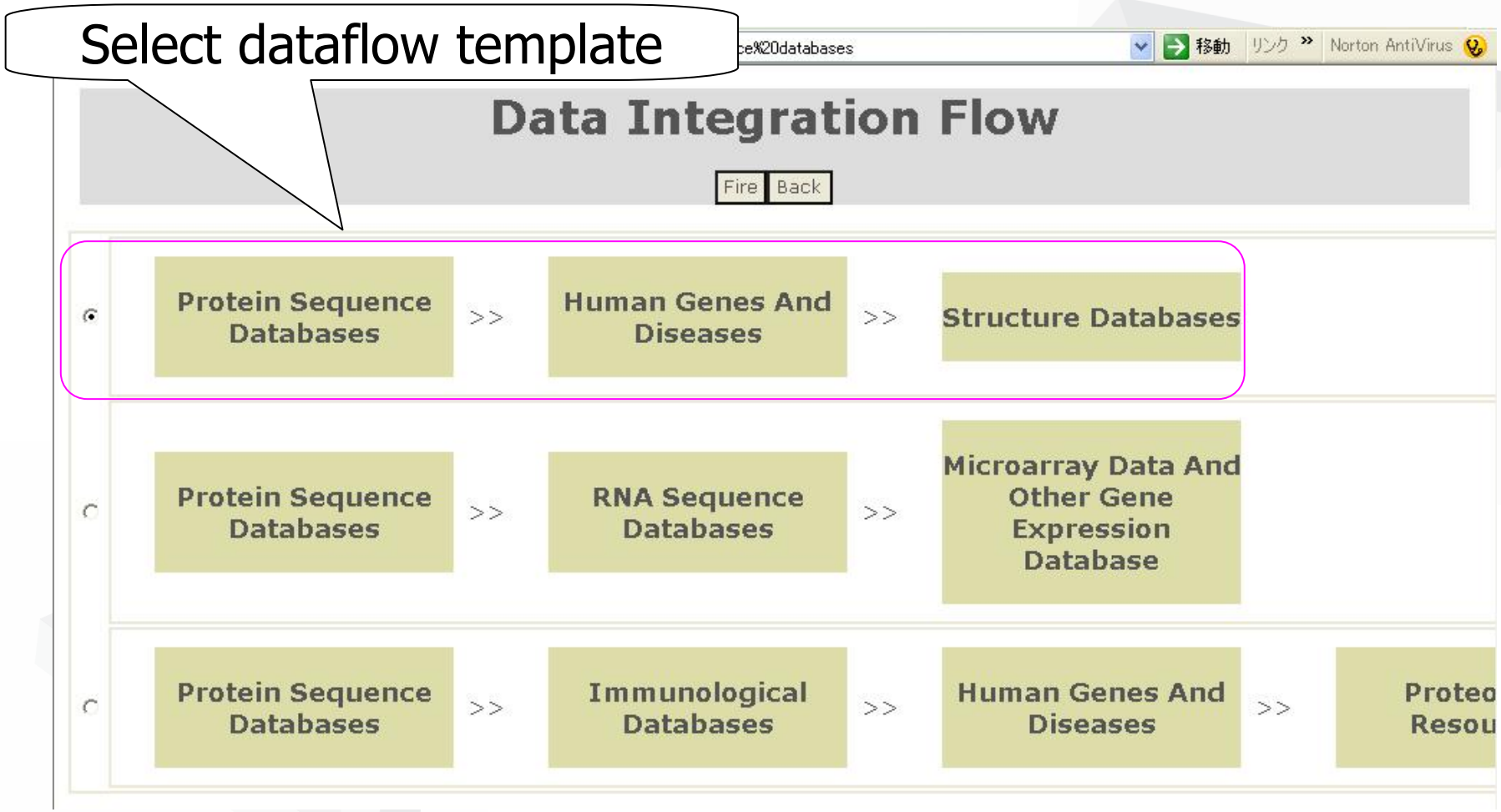
Keyword Search 1



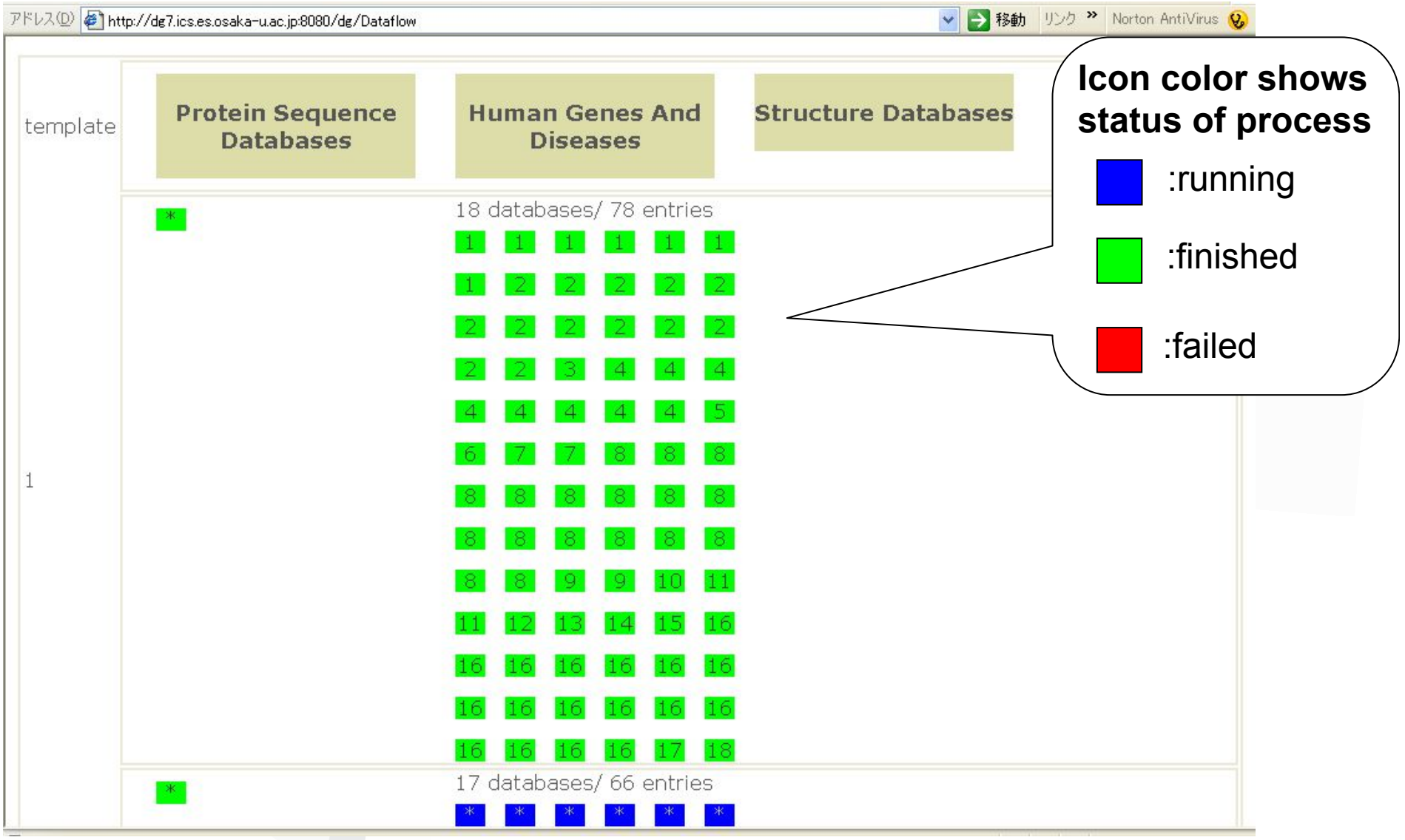
Keyword Search 2



Data Integration Flow 1



Data Integration Flow 2



Summary and Future Works

- We have developed a system for data integration of >100 lifescience DBs.
- Globus Toolkit and OGSA-DAI integrates distributed DBs and metadata.
- DB keywords and indices are stored in a grid filesystem using Gfarm.

Future Works

- Data are currently downloaded. Need to use web-service (WSDL/SOAP) interface.
- Data-intensive workflow tool is still under development.