



# Collating Data for Ease of Browsing and Exploration

Exploit your data for better decision-making

Lewis Jardine, Intellidos



# Outline

- Investment in Data
- Current Visibility of Data
- Drive to Data Integration
- Collation by Entity
- Questions
- Querying Integrated Data
- Query by Entity
- Some Solutions

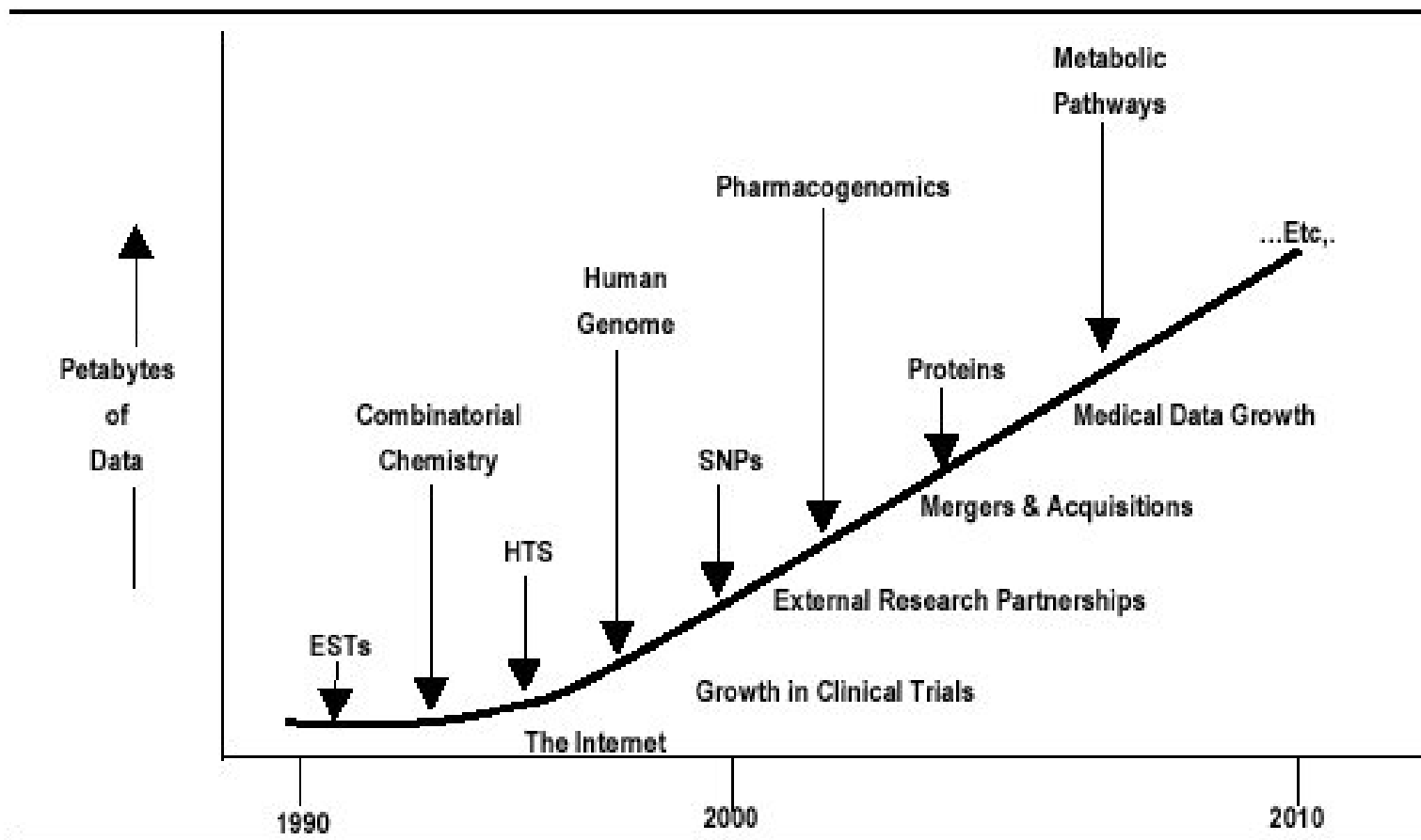


# Investment in Data

- Pharmaceutical and biotechnology companies spend billions of dollars a year assembling research databases.
- Technologies such as combi-chem, HTS, genomics, expression profiling, pharmacogenomics, and proteomics have lead to vast increases in both the volume and number of types of data.
- Resulting in petabytes of heterogeneous data!



# Technologies & Data Growth



Source: IBM Life Sciences



# Purpose of Data

- Provide Information
- Enhance knowledge.
- Enable decision making.
- Develop New Drugs!

**If this is not achieved  
it invalidates the business case  
for investing in data!**

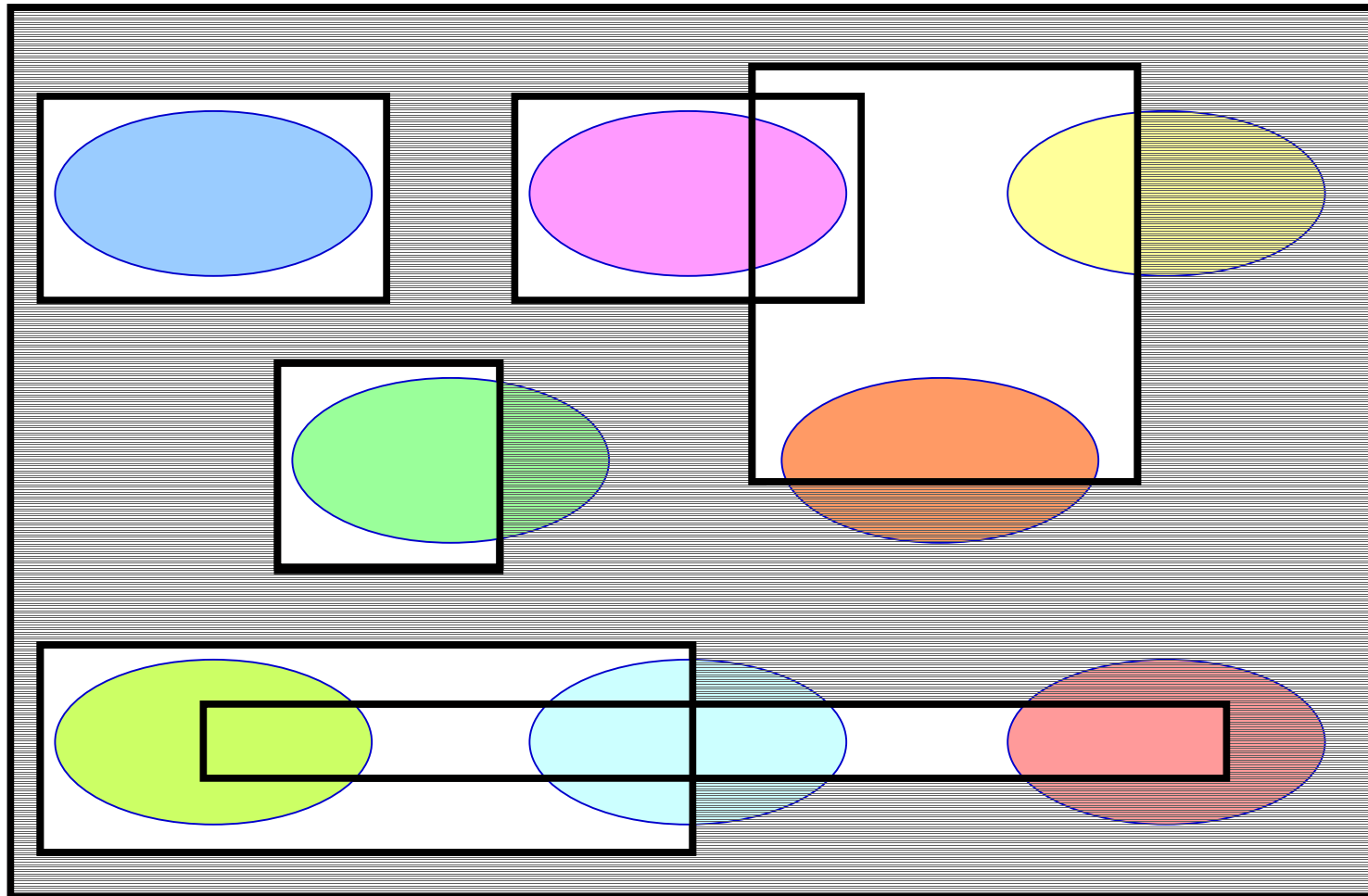


# User Profile

- Users will not have extensive IT skills.
- They will range from scientists to managers, statisticians, administrators...
- All will have varied overlapping needs.
  
- Note that this does not just apply to Discovery but all the way along the drug development pipeline through to clinical trials and possibly beyond.



# Current Visibility of Data



Restricted windows or views of the data.

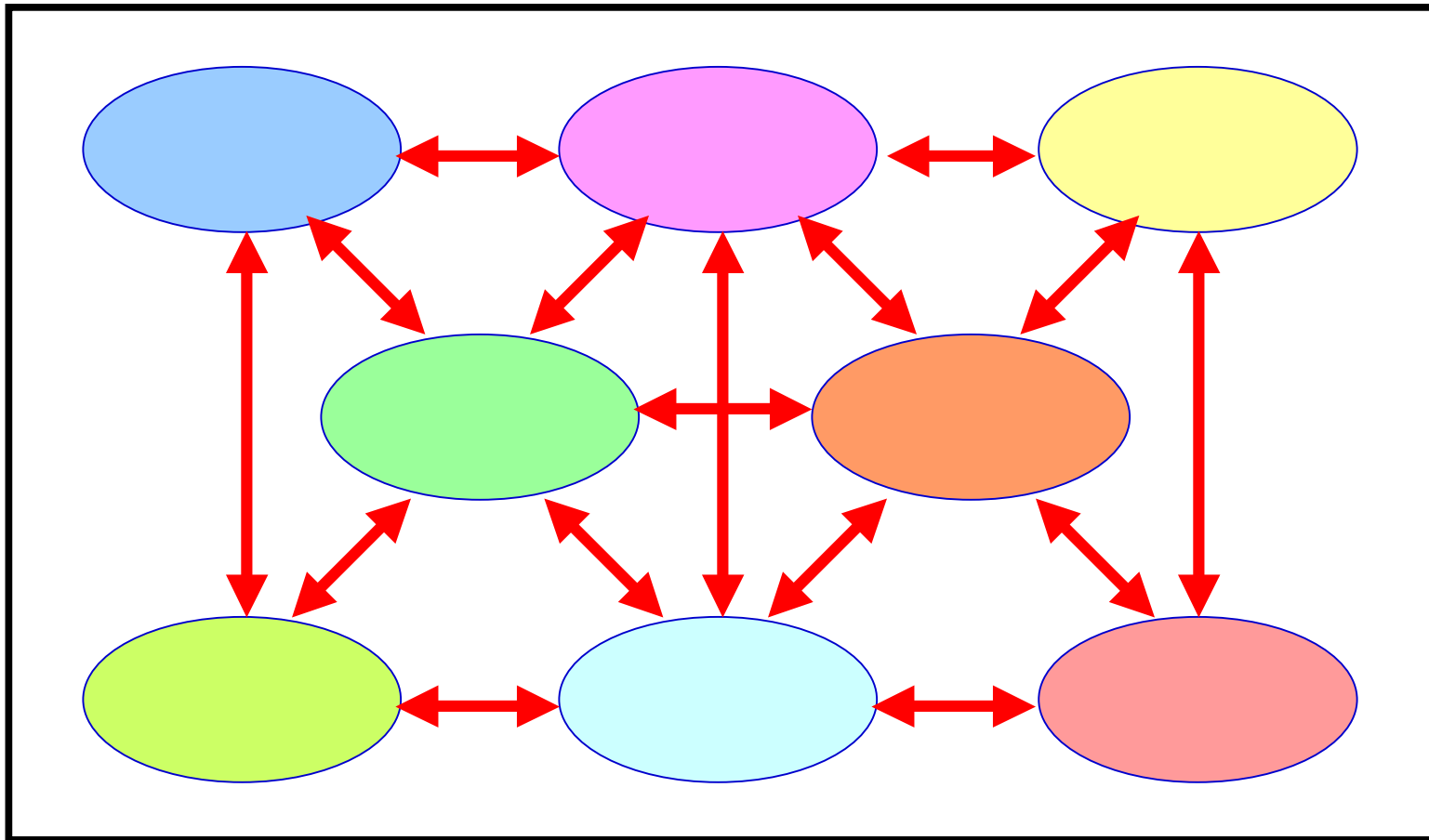


# Current Visibility of Data

- Typified by a number of specialized applications viewing each data source.
- A few make use of a small number of different sources of data.
- Fewer still look at many data sources and these typically through a small window.
- Very little transfer of knowledge up and down the drug development pipeline.



# Potential Value of Data



View of all the data connected together.

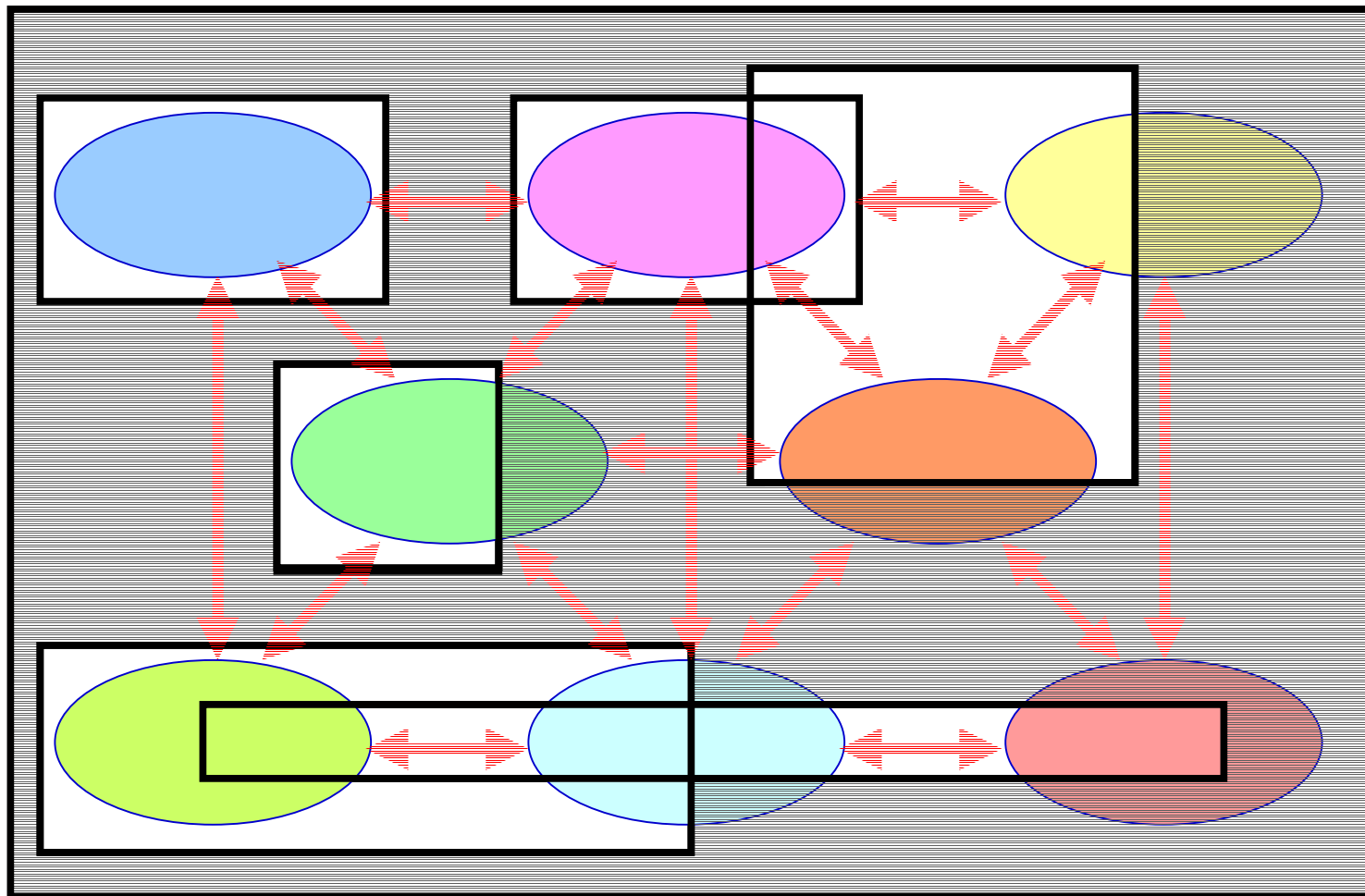


# Drive to Integration

- The benefits of data integration have resulted in a massive push towards federated systems or warehouses.
- Considerable effort is being put into back-end data integration.
- However, less effort is being put into allowing users to interact with integrated data.



# Current Effect of Integration



Integration infrastructure brings little benefit to normal users.



# Downside to Data Integration

- More data !
- More schemas !
- More joins !
- Offers multiple possible joins.
- Requires more complex queries.
- Lack of centralized ownership.
- The knowledge of databases and schemas will remain distributed.



# Collation

- One way to reduce the complexity is to approach the problem by collating by entity type rather than database.
- This has two major advantages:
  - There are fewer entity types than DB tables
  - Users understand entity types better than databases.
- Additionally, it supports the distribution of knowledge of domains & schemas.

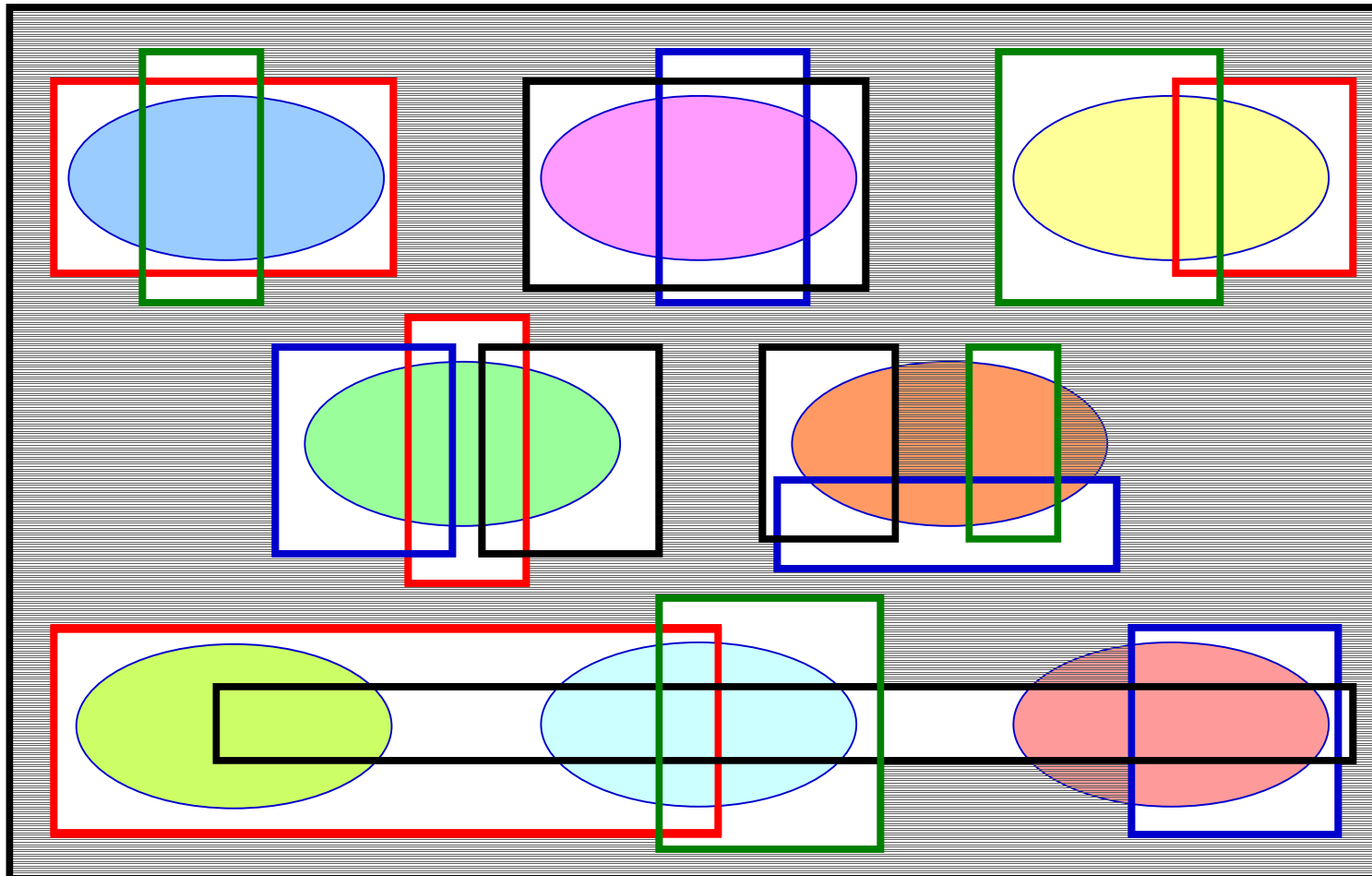


# Example Entity Types

- Batch / Lot
  - Compound
  - Assay
  - Protein
  - Protein Family
  - Gene
  - Person
  - Team
  - Project
  - Task
  - Skill
  - Company
- Also provide annotation functionality to allow users to comment on or register interest in specific entities.



# Data Collation by Entity Type



Views are coloured by collation and referenced via the entity ID.



# Example Collation GUI

Small Molecule:

- Registration Details
- 2D Structure
- 3D Structure
- In-house Assays Data
- NCI Assay Data
- NCI Properties
- Hazards
- Annotations
- Tautomers (Engine)
- Similarity (Engine)
- Batch Availability

3 assays were run for this compound

ID	Assay	Activity	Range	Units
<a href="#">ID-71M001-001</a>	Locomotor Response	6246.37	398.12	nMol
<a href="#">ID-71M001-001</a>	Locomotor Response	233.44	15.3	nMol
<a href="#">ID-71M001-001</a>	Serotonergic appetite control	4726.3	301.34	uMol

Export options: [CSV](#) | [MS Excel](#) | [XML](#)

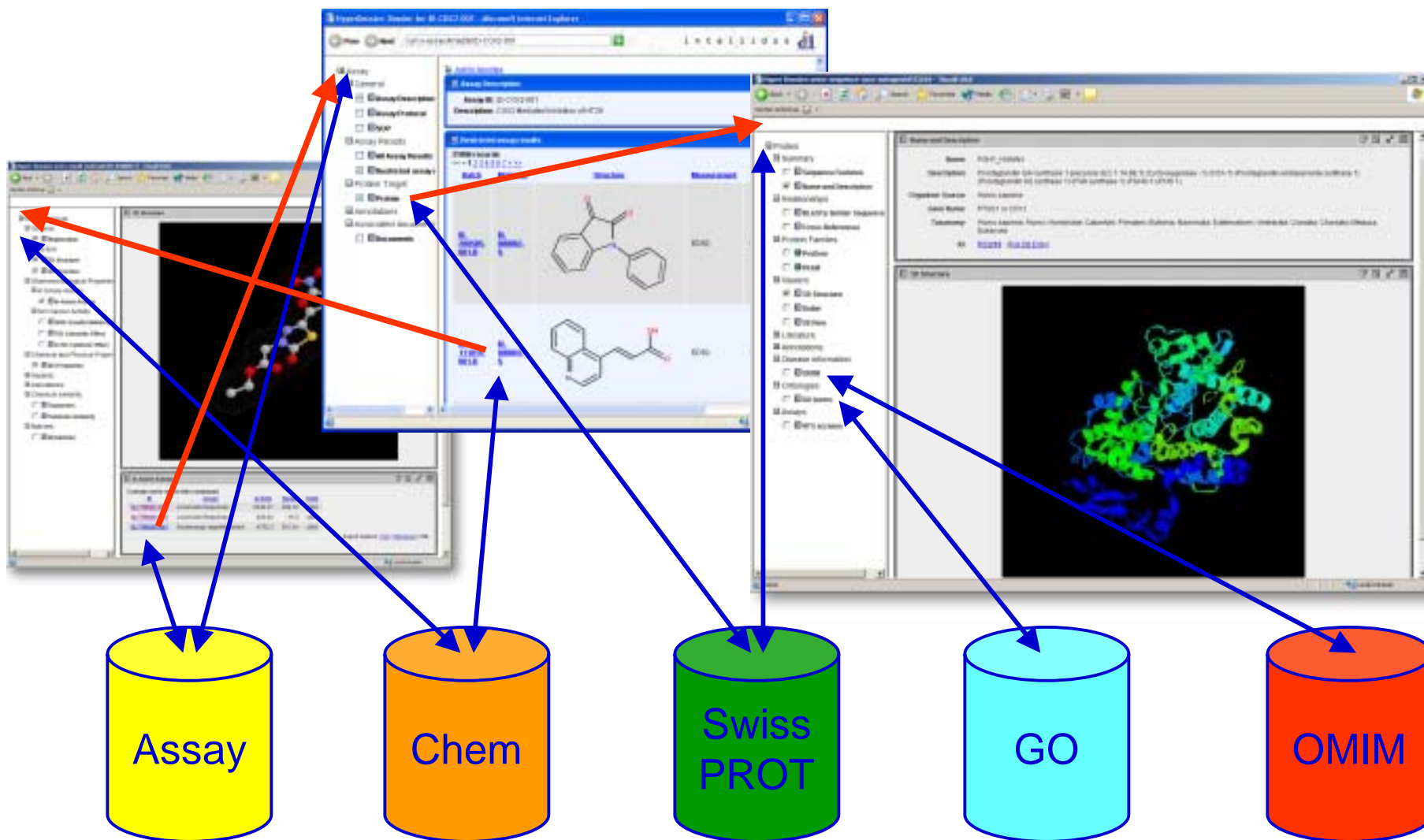


# Linking Collated Data Sets

- The overlap between collated data sets is a significant benefit.
- This overlap can be used to define links between data in a collated data set and another entity and its collated data set.
- These can be expressed as hyperlinks.
- Allowing browsing and exploring of collated data.



# Linked Collations





# Benefits of Collation

- Provides an 'entity-centric' view.
- Easy for users to understand.
- Hides the complexity of DBs & schemas.
- Supports integration of information via linking.
- Enables browsing and exploration.
- Can support distributed data ownership if good administration tools are provided.



# Weaknesses of Collation

- May require high set up costs.
- May require significant administration.
- Security issues caused by accessing data from all sources must be handled.
- Does not offer any solution to the problems of querying integrated data.
  - Typical questions follow...



## Queries

- Scientists need to be able to ask complex cross-domain questions.

“Show me all the compounds that have been tested against members of the serotonin family of receptors, have IC50 values in the nanomolar range, a molecular weight between 200 and 500, and a logP between 3 and 5.”



## Operational Questions

- To set up a ‘clean’ project team with no connection with an old alliance which covered the same target area.

“Give me all the people outside the old alliance project team with skills in chiral synthesis and who have not looked at any targets generated by that alliance?”



# Business Intelligence Example

- To select areas for possible research.

“Show me all the genes which have a known association with cancer, lie on chromosomes 2 or 5, have a mouse knockout and none of our competitors have a compound for this target in Phase I clinical trials or later.”



# Querying Integrated Data

- Traditional approaches such as forms do not scale well to integrated data:
  - One huge extremely complex form.
  - Very many smaller forms.
- Even DBAs have problems due to the number and complexity of schemas.
- Joins become especially problematic.
  - Many possible joins between two tables but not all are scientifically valid...



# An Entity Based Approach

- The creation of collations builds up a set of relationships between database content and domain entities.
- These can be reversed to form queries.
- The creation of links between related collations builds up a set of 'valid' relationships between these entities.
- These describe joins between the data sources.



# Query by Entity

- Provided all the information required to create a set of collated datasets is captured. It should be possible to reverse and massage this 'metadata' to create an entity-centric query engine.
- Operates in a scientific context space.
- Complete abstraction from schemas.
- Much easier for users to understand.
- Administration is already done.



# Potential Benefits

- Allow scientists and other decision makers to query in their domain context.
- Enable effective access and utilisation of all the available data.
- Reduce the duplication of effort.
- Save on administration effort.
- Improve decision making.
- Improve the efficiency of drug discovery.



# About Intellidos

- Intellidos is a team of highly creative software engineers and scientists who provide innovative solutions tailored to the most challenging needs of our customers.
- Our main products are our flexible ad-hoc query tool **QueryConstructor™** and the **HyperDossier™** data browser.



# HyperDossier



### Search Wizard

Select a search strategy:

- My Stored Queries
- Global Public Stored Queries
- QueryConstructor™
- Full Text
- BLAST
- Chemical Substructure
- Chemical Similarity

### Recent Dossiers

ID	Last
urn:small-molecule:ID-000003-S	1 Apr 12 11
urn:assayarrayDB:ID-COX2-001	1 Apr 12 08
urn:small-molecule:ID-004190-S	17 Mar 17 20
urn:pubchem.ncbi:pubchem:P25264	17 Mar 17 20
urn:pubchem:CEAN00001	17 Mar 17 19

### My Favorites

Name	Date
AssayArray	5 Dec 2003 16:47:33
Leads	5 Dec 2003 16:39:29
COX2 Assay	5 Dec 2003 16:14:37
ESIC: 000001	5 Dec 2003 16:04:09

### Direct Search

urn:small-molecule:ID-000003-S

### HyperDossier: Dossier for ID-COX2-001 - Assay

- Assay
  - General
    - Assay Description
    - Assay Protocol
    - SOP
  - Assay Results
    - All Assay Results
    - Restricted assays
  - Protein Target
    - Proteins
  - Annotations
  - Associated documents
    - Documents

### HyperDossier: Dossier for ID-000003-S - Microsoft Internet Explorer

Intellidos

- Small\_Molecule
  - General
    - Registration
  - Structure
    - 2D Structure
    - 3D Structure
  - Observed Biological Properties
    - In House Assays
    - In House Assays
    - NO Cancer Activity
    - K1616 Growth Inhibitory Power
    - TIG Cytotoxic Effect
    - C53 Cytotoxic Effect
  - Chemical and Physical Properties
    - NO Properties
  - Hazards
  - Annotations
  - Chemical similarity
    - Frameworks
    - Framework similarity
  - Batcher
    - All batches
  - Metabolic Property
    - Alliance associations
    - Logged accesses

### Assay Dossier

English view

NSC	180026
CAS registration number	17024-20-5
Molecular Weight	180.20080
Formula	C12H10N2O2

### 3D Structure

### Restricted assays

31000 records

ID	ID	EC50	ED50	ED50
urn:pubchem:CEAN00001	urn:pubchem:CEAN00001		ED50	5823.37 371.10
urn:pubchem:CEAN00001	urn:pubchem:CEAN00001		ED50	7183.53 458.51



# Administration

The screenshot displays the Intellicol HyperDossier administration interface. On the left, a sidebar lists management options: Users, Dossiers and Reports, Alerts, Data Sources, Providers, Logs, and System Information. The main content area shows a table of dossier types with columns for Dossier, Description, Entity Type, Default, Created, and Actions.

Dossier	Description	Entity Type	Default	Created	Actions
Person	Information about a person	x-user		13 Mar 2003 00:53:50	Edit Save Delete
Alliance	A dossier about alliances between companies	x-alliance			
Budget	This dossier contains information about a given budget	x-budget			
Location	This dossier contains information about a given location	x-location			
Project	This dossier contains information about a given project	x-project			
Skill	A dossier about a particular skill	x-skill			
Task	This dossier contains information about a given task	x-task			
Team	Information about a team	x-team			
Company	A dossier about a company or institution with which there is a business relationship	x-company			
Document	A document that may be an external publication or an internal document which may have one or more versions	x-document			
Disease	A description of a disease or syndrome taken from a particular (Skill) entry	x-disease			

An SQL query window is open, showing the following query:

```
SELECT  
a.MRNM AS AB_Sing_ID,  
a.FIRM_name AS Firm_Name AS Name,  
a.Ant_Site AS Ant_Site,  
c.Company_id AS Company_id,  
c.Company_name AS Company_name,  
a1.Contact_address AS Contact_Address,  
a1.SM_code AS SM_code,  
a1.SM_name AS SM_name,  
a1.Telephone AS Telephone,  
a1.Phone_number AS Phone  
FROM
```

The results table below shows employee data:

EmpID	Name	Job Title	Company ID	Company Name	Email Address	Site Code	Site	Time Zone	Phone
EMP0001	Andrew Park	Chief Executive Officer	COMP0001	Intellicol Ltd	andrew.park@intellicol.com	STP0001	Office	0	+44 (0) 1235 838 002
EMP0002	Lewis Jardine	Engineering Director	COMP0001	Intellicol Ltd	lewis.jardine@intellicol.com	STP0001	Office	0	+44 (0) 1235 838 003
EMP0003	Steven Parker	Operations Director	COMP0001	Intellicol Ltd	steven.parker@intellicol.com	STP0001	Office	0	+44 (0) 1235 838 004





# Intellidos

- Intellidos also offers consulting and bespoke software development...

[lewis@intellidos.com](mailto:lewis@intellidos.com)

[www.intellidos.com](http://www.intellidos.com)