MERCK SERONO

# Integration of in-house and external data in practice

D. Domine, C. Merlot, M. Ibberson, and M. de Francesco
Drug Discovery Informatics
Merck Serono Geneva Research Center

# Outline

- Introduction: Drivers and Needs in KM

- Merck Serono Knowledge Space
  - In-house data integration
  - External data integration
  - Workflows
  - Project management
  - Entity aggregation

- Examples
- Concluding remarks

# Drivers

- **Informed decision-making at different levels: Fast, easy, well-informed.**
- **Increased complexity, variety, quantity of information/technologies**
- **Broader scope (e.g Discovery, Clinics, Manufacturing, etc)**
- **More data outside than inside. Need better integration with in-house information**
- **All project team members and management need to understand the global status and goal**
- **Navigation between different levels of abstraction.**
- **Know-how capitalization**
- **Optimize resources**
- **Enable mergers**

# Needs

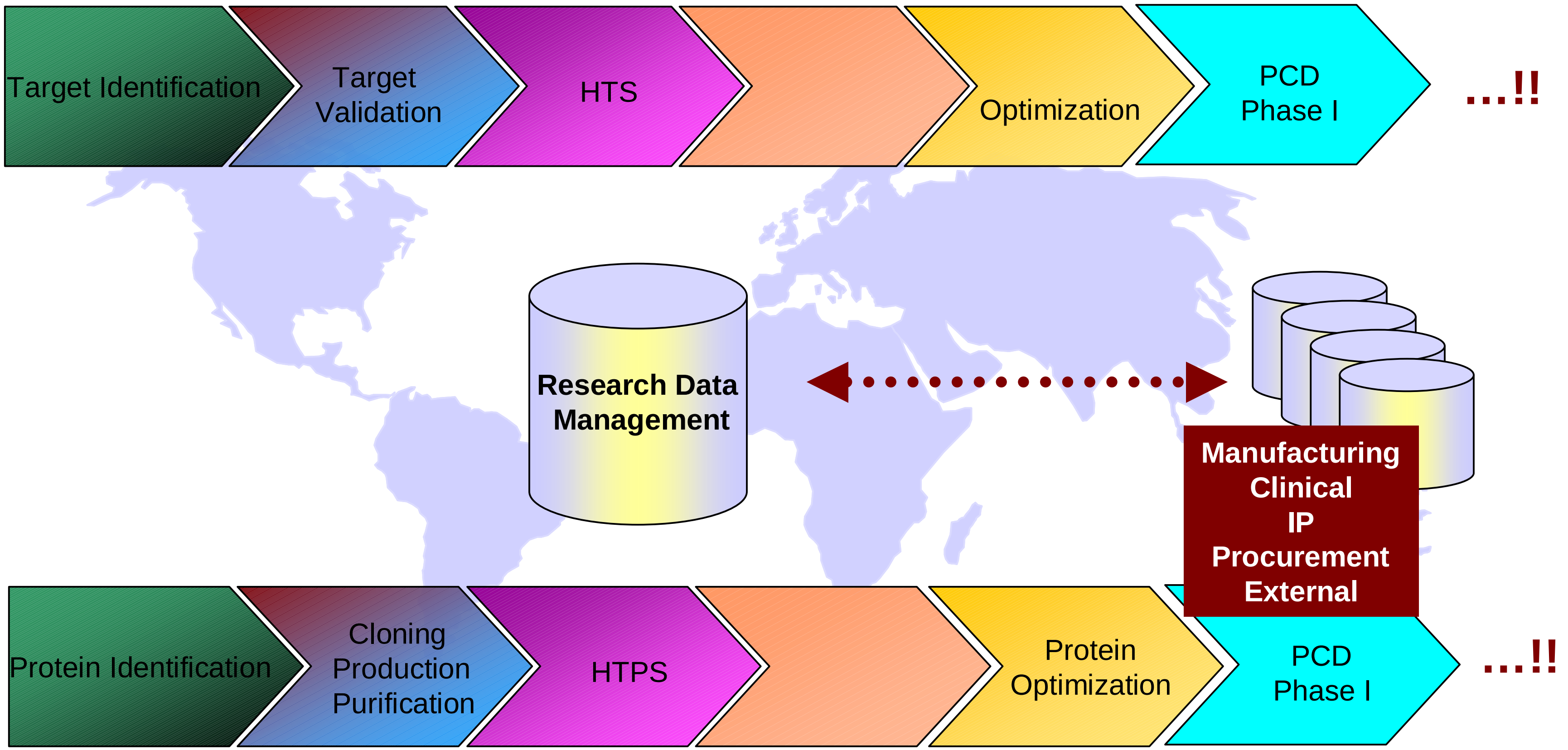
- **In-house experimental data integration. Never ending**
  - Constantly evolving
  - Broader scope (Chem., bio., pharm., clin., gen., manuf., IP, Compet., ….)
  - Capture interpretations vs experimental data, results
- **External data/Prediction tools integration. Starting**
  - Need to be structured
  - Requires manual curation, understanding
  - Lack of standards
- **Know-how, unstructured data acquisition, mining, interpretation (in-house and external). Starting**
- **Data and Information Mapping/Scoring. Starting**
- **Different levels of understanding/reporting (lab., project, Management). Starting**
- **Automated annotation/interpretation to turn data into information and knowledge Possible at all?**

# Some existing technical solutions

- **Integration: Function of needs and company history**
  - Requisite for automated interpretation
  - Links – Cross-references
  - Indices (sequence dbs, DiscoveryGate).
  - Metalayers (description of underlying data sources and business rules)
  - Data marts (duplication of information)
  - Standards: Unique definition of entities and relations (e.g. Semantic web)
- **Know-how acquisition**
  - ELNs
  - Documentation management
- **Unstructured data mining**
  - Text mining
- **Information mapping/scoring**
  - Mostly scoring
  - Viewers (e.g. sequence viewers)

# Research Context

## Small molecules


Target Identification

Target Validation

HTS

Optimization

PCD Phase I

…!!

Research Data Management

Manufacturing
Clinical
IP
Procurement
External

Protein Identification

Cloning Production Purification

HTPS

Protein Optimization

PCD Phase I

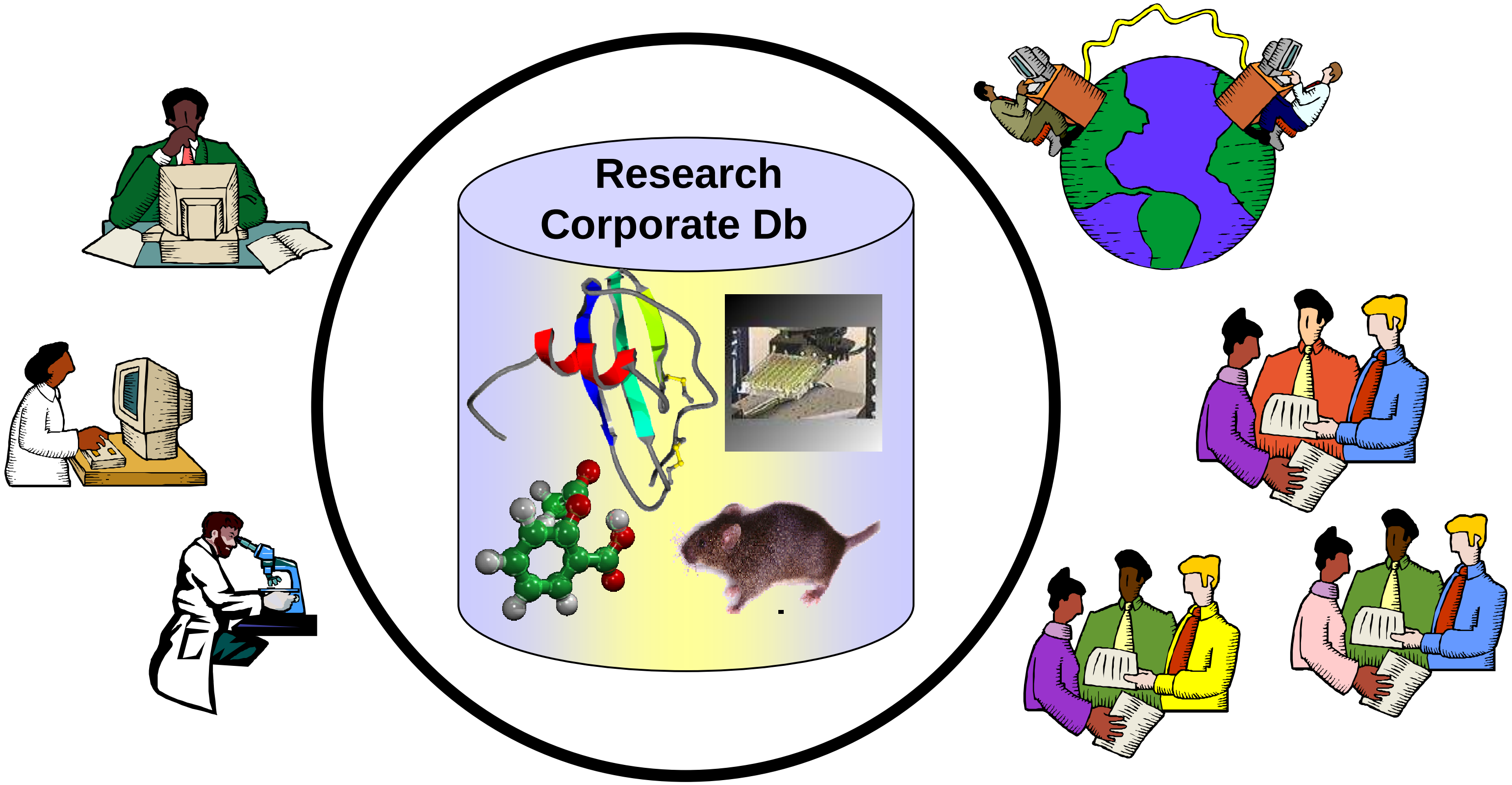
…!!

## Proteins and Antibodies

**Right (Right) information to the right people at the right time in the right context**

# A Culture of Sharing

- **One place** for all biological, chemical & pharmacological information
- No boundaries **between sites, teams, projects**



Research Corporate Db

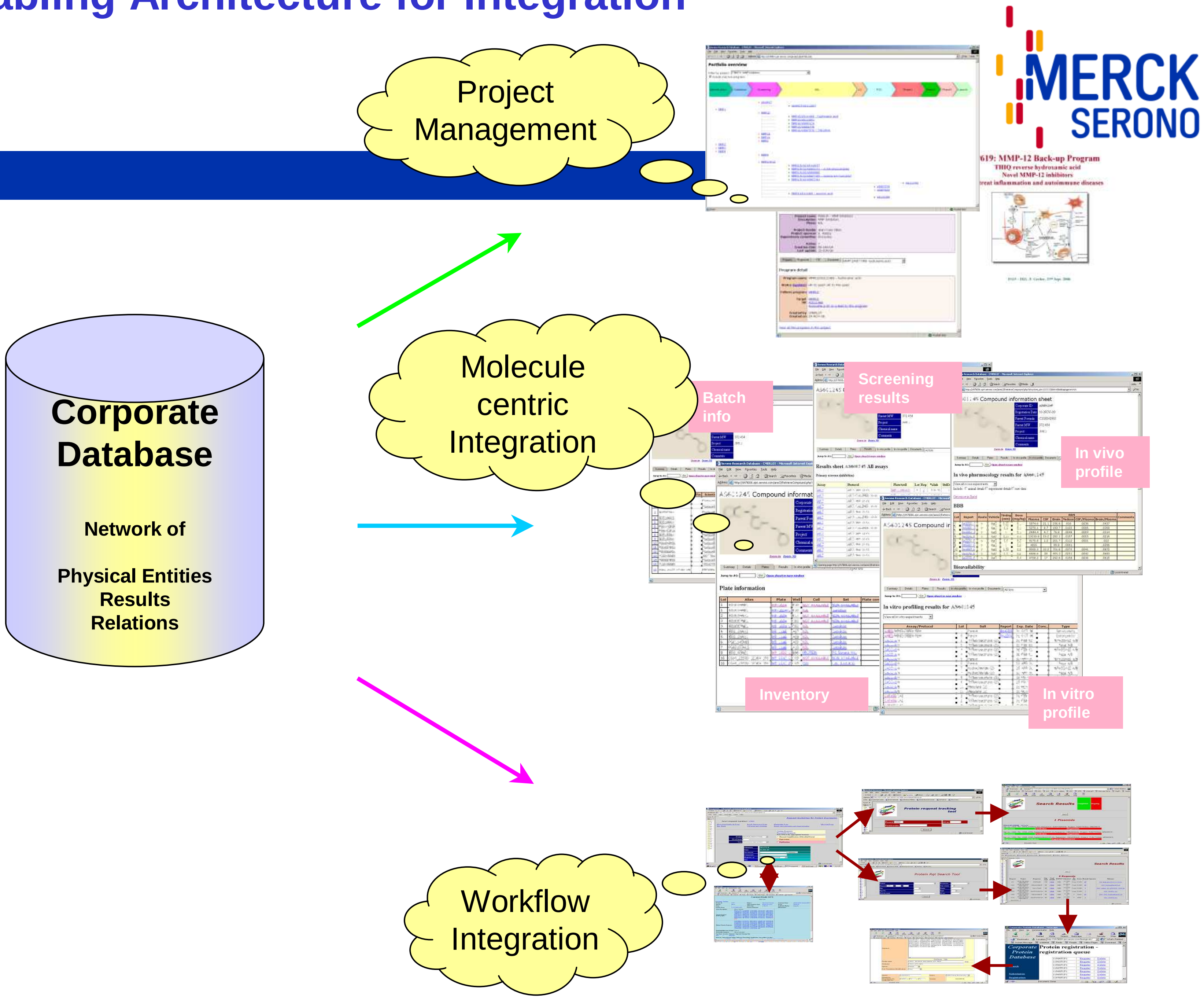
**Access to well integrated information:**

**- to make people “knowledgeable”**

**- to facilitate information flows**

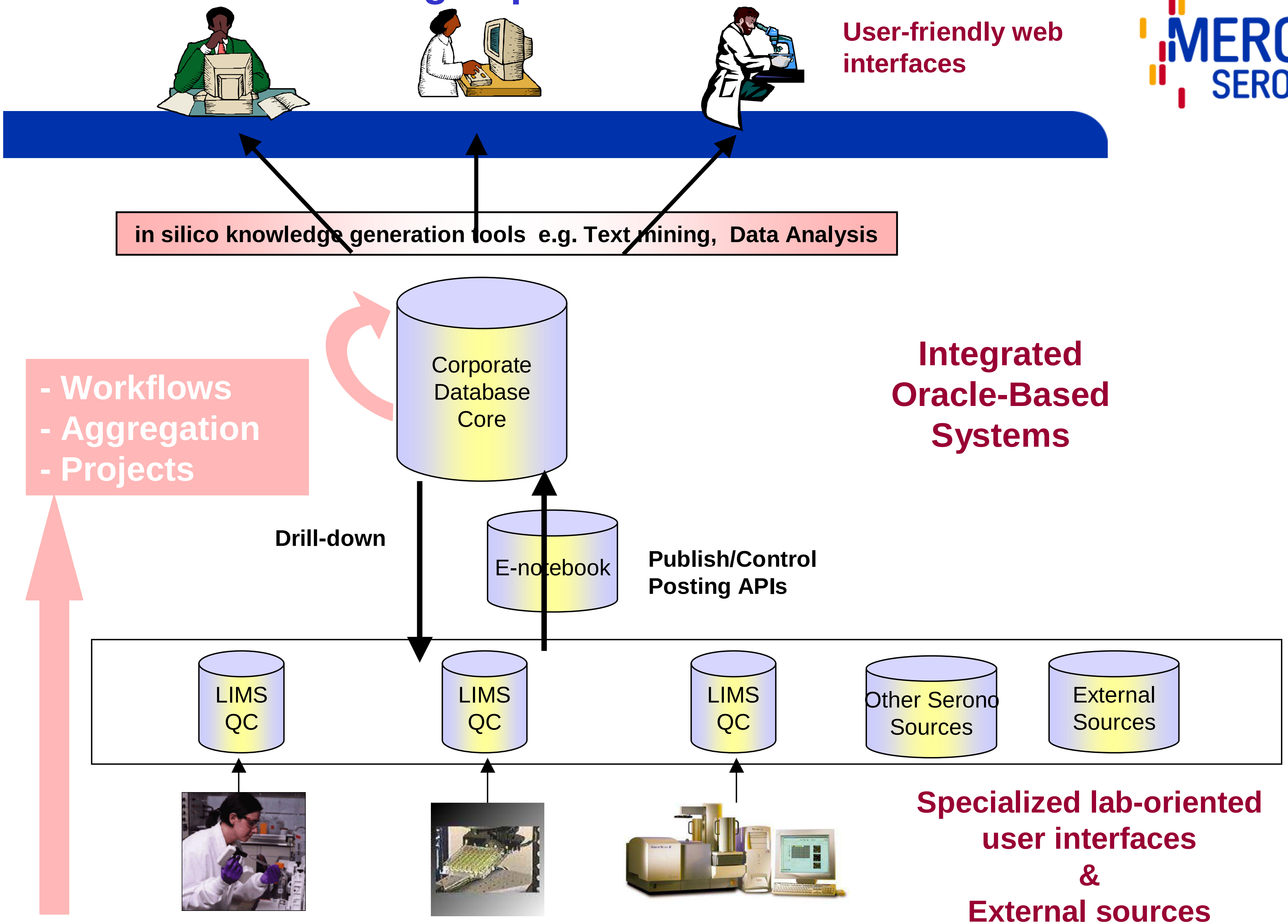
**- to facilitate information tracking**

# An Enabling Architecture for Integration



Corporate Database
Network of
Physical Entities
Results
Relations
Project Management
Molecule centric Integration
Workflow Integration
Batch info
Screening results
In vivo profile
Inventory
In vitro profile
MERCK SERONO
619: MMP-12 Back-up Program
THIQ reverse hydroxamic acid
Novel MMP-12 inhibitors
treat inflammation and autoimmune diseases
Portfolio overview
Results sheet
In vivo pharmacology results
Plate information
In vitro profiling results
Protein request tracking tool
Search Results
Corporate Protein Database
Protein registration - registration queue

# Research Knowledge Space - Framework



User-friendly web interfaces

in silico knowledge generation tools e.g. Text mining, Data Analysis

Corporate Database Core

Integrated Oracle-Based Systems

- Workflows
- Aggregation
- Projects

Drill-down

E-notebook

Publish/Control Posting APIs

LIMS QC

LIMS QC

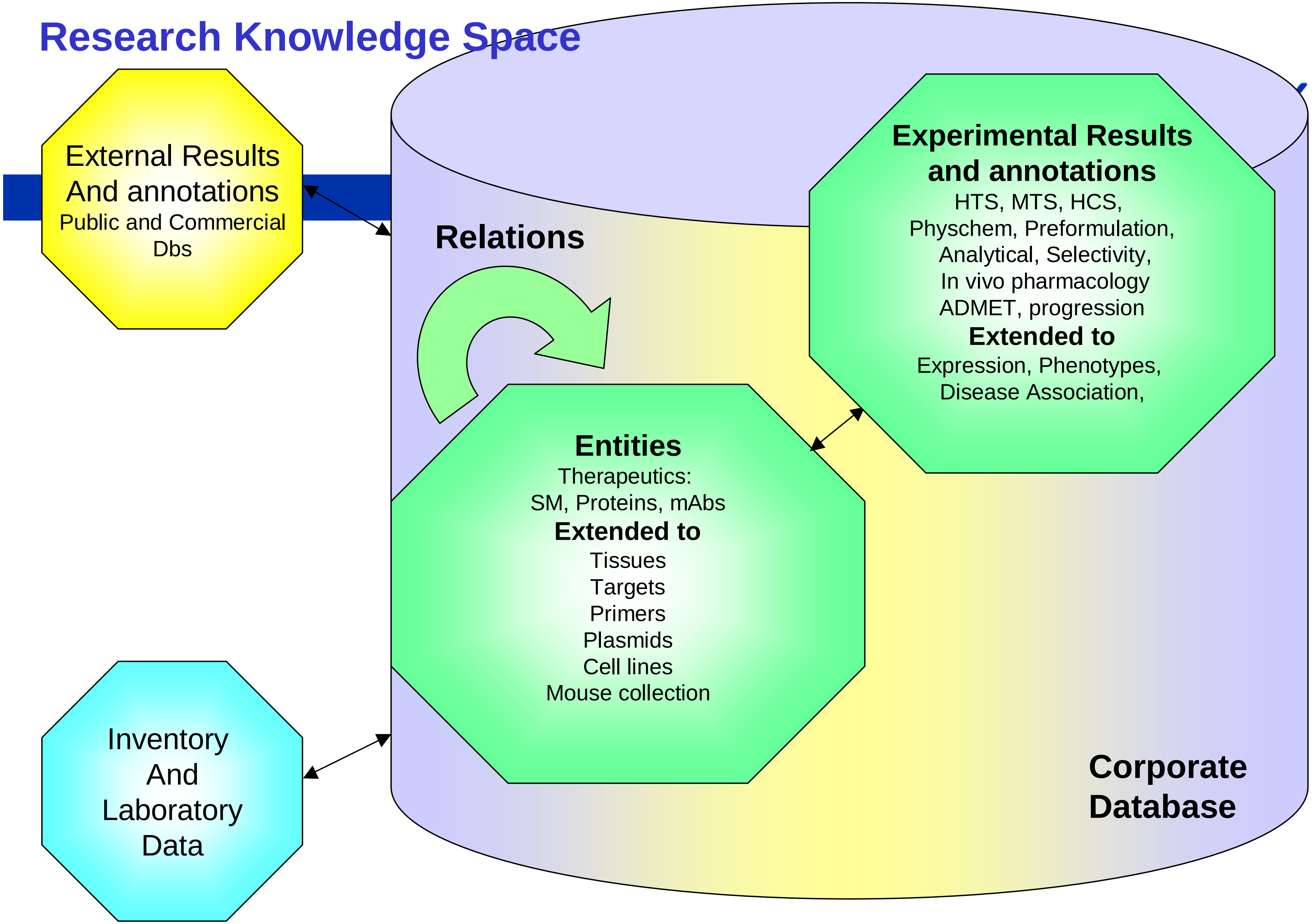
LIMS QC

Other Serono Sources

External Sources

Specialized lab-oriented user interfaces & External sources

# Research Knowledge Space



External Results
And annotations
Public and Commercial
Dbs

Relations

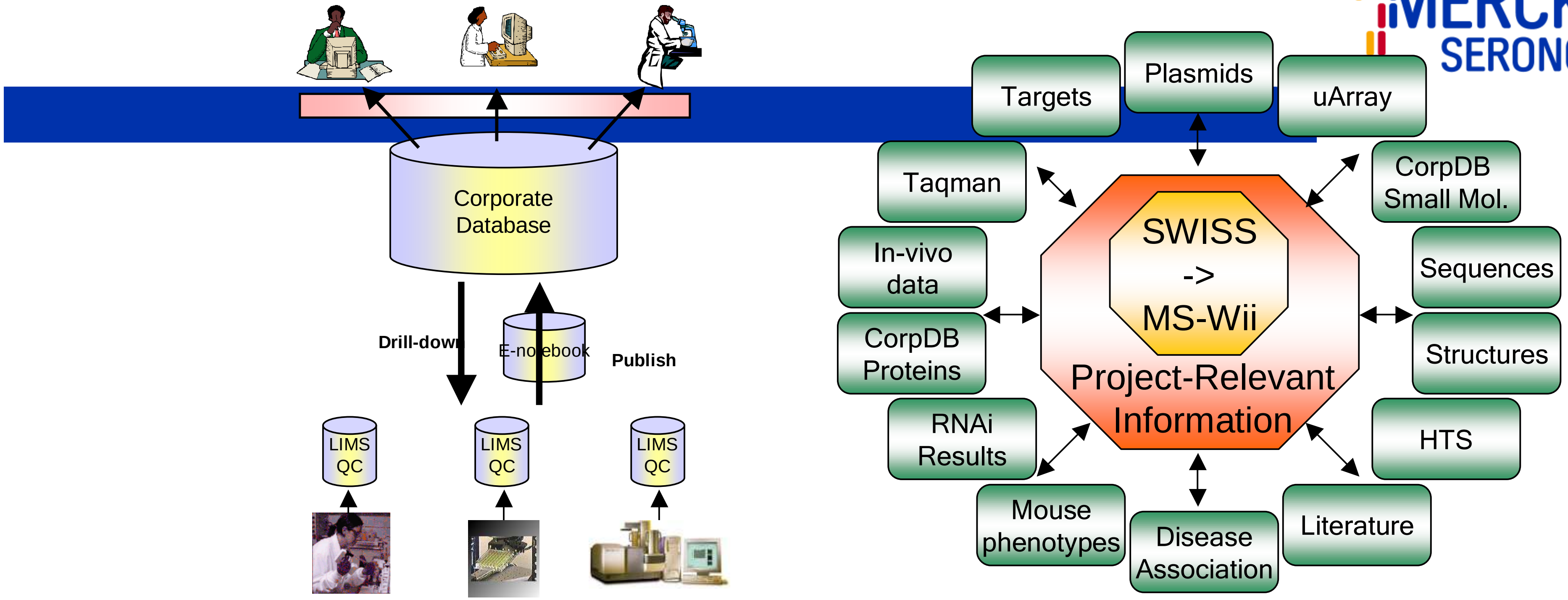
Experimental Results
and annotations
HTS, MTS, HCS,
Physchem, Preformulation,
Analytical, Selectivity,
In vivo pharmacology
ADMET, progression
**Extended to**
Expression, Phenotypes,
Disease Association,

**Entities**
Therapeutics:
SM, Proteins, mAbs
**Extended to**
Tissues
Targets
Primers
Plasmids
Cell lines
Mouse collection

Inventory
And
Laboratory
Data

**Corporate
Database**

# Knowledge Space – Framework – Molecule centric



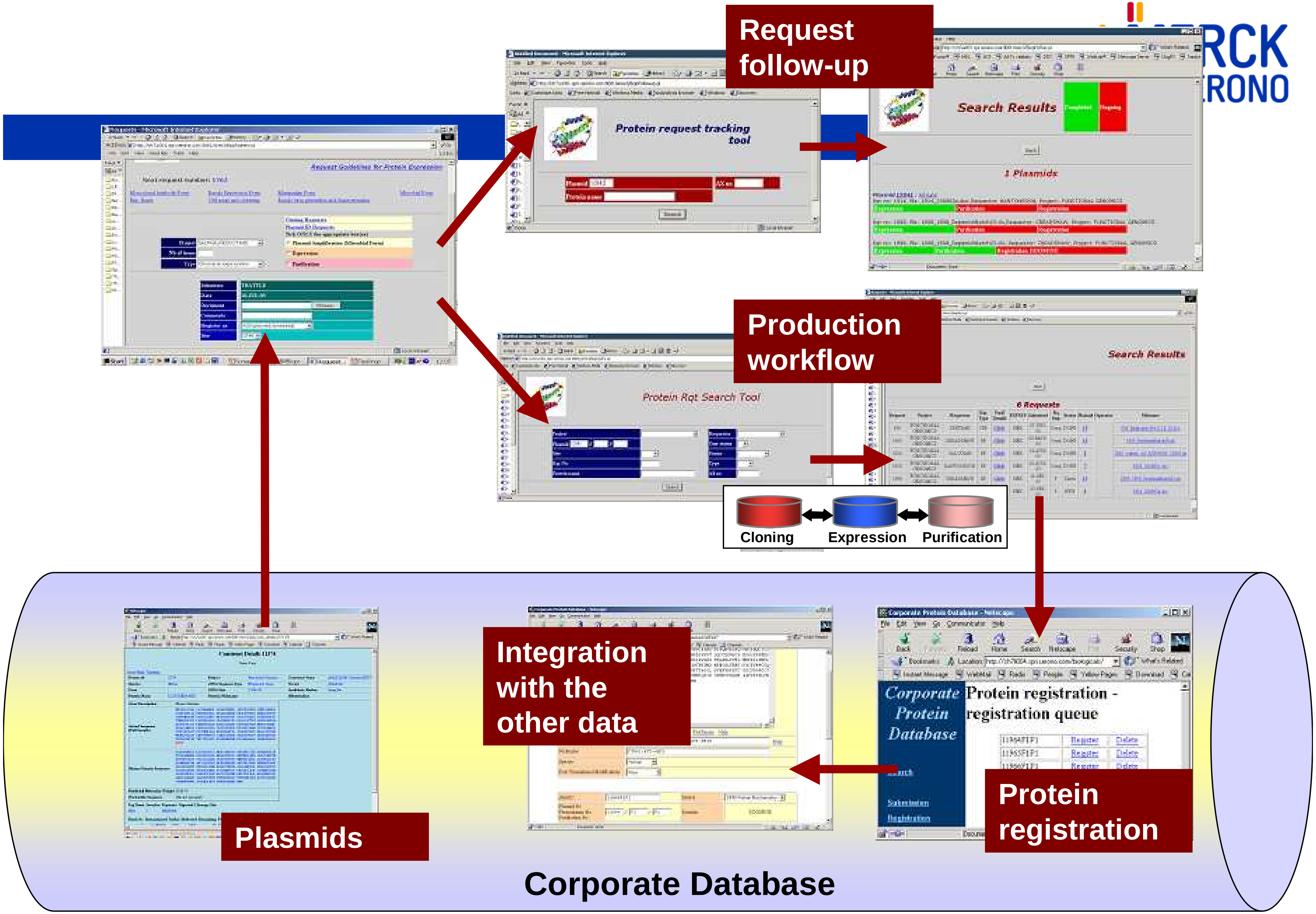
**System comprising of:**

- **a central gateway for handling entities (e.g.: genes, proteins, chemicals): the Merck Serono Wide Index (MS-Wii)**
- **specialized databases to handle large amount of raw data, each specific of a technology, referring to entities in the SWI**
- **a system to manage project-relevant information. Also, an easy-to-understand entry point for more detailed results and raw data (EDS)**
- **a tracking of production workflows**
- **same system for external data sources -> One single entry point**

# Knowledge Space – External

- **Several categories:**
  - low-throughput: Must be curated, but high value. eg.: NTP reports database, public FDA databases
  - high-throughput: Large databases integrating heterogeneous data. Public large scale experiments e.g.: ENSEMBL, MDL MDDR, McKay
- **Integration**
  - Index in MS-Wii
  - Filter information relevant to the company's objectives.
- **Some issues:**
  - Some servers might not accept to full download
  - Need to build an appropriate analysis pipeline to interpret raw data
  - Requires prior analysis of structure and contents to design parsing scripts
    - "Thought-to-be structured"
- **Examples:**
  - 1- Sequences: new releases parsed and integrated in MS-Wii.
    - 1 s to get all in-house and external information for a given sequence whether it is a gene, transcript or protein
    - Information on past releases kept
    - Initiated same work on small molecules
  - 2- Toxicity data and predictions.

# Workflow Integration: Bioprocessing Pipeline



Request Guidelines for Protein Expression
Protein request tracking tool
Search Results
1 Plasmids
Protein Rqt Search Tool
Search Results
6 Requests
Corporate Protein Database
Protein registration - registration queue

Request follow-up

Production workflow

Cloning ⟷ Expression ⟷ Purification

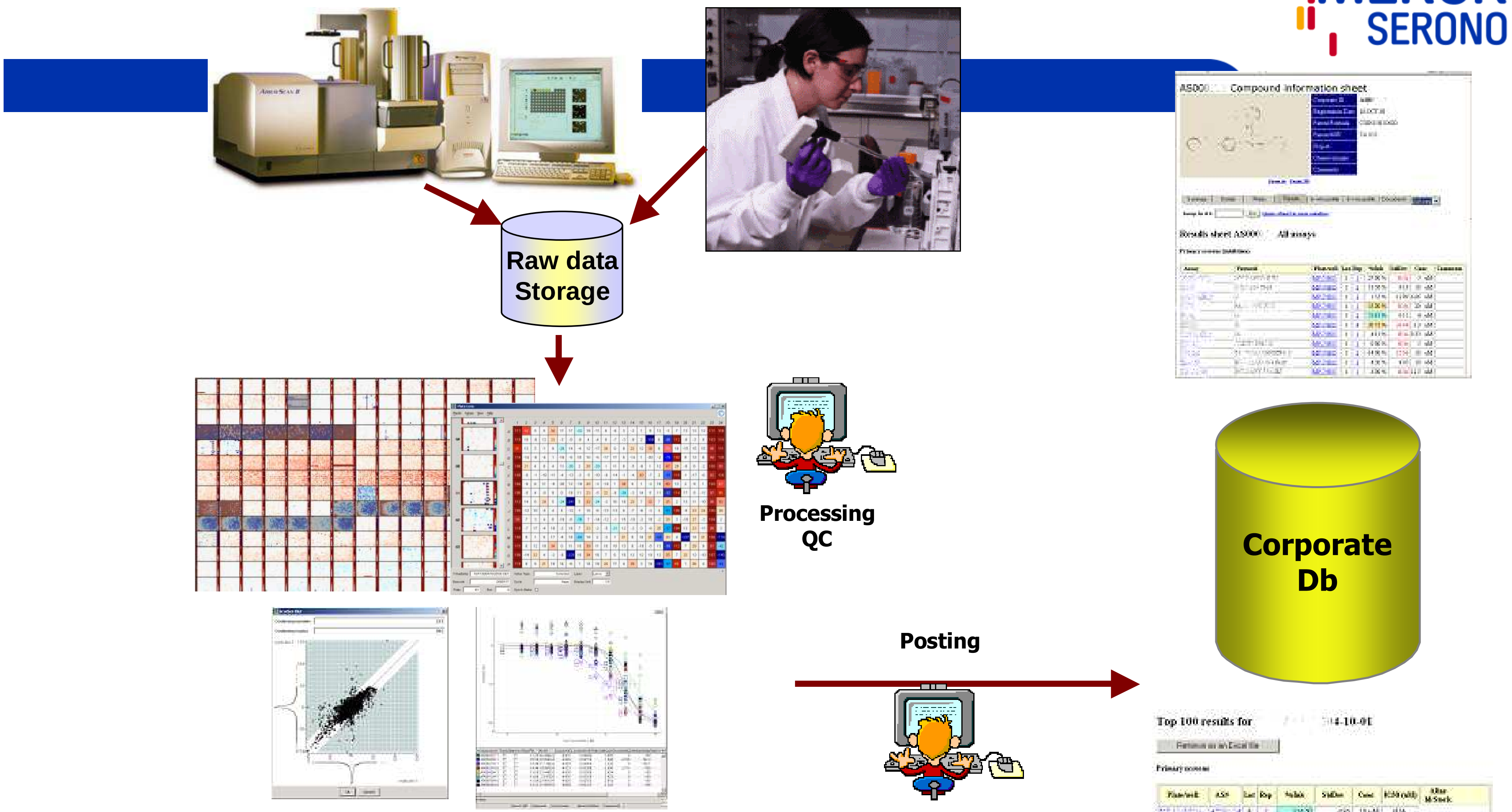
Integration with the other data

Plasmids

Protein registration

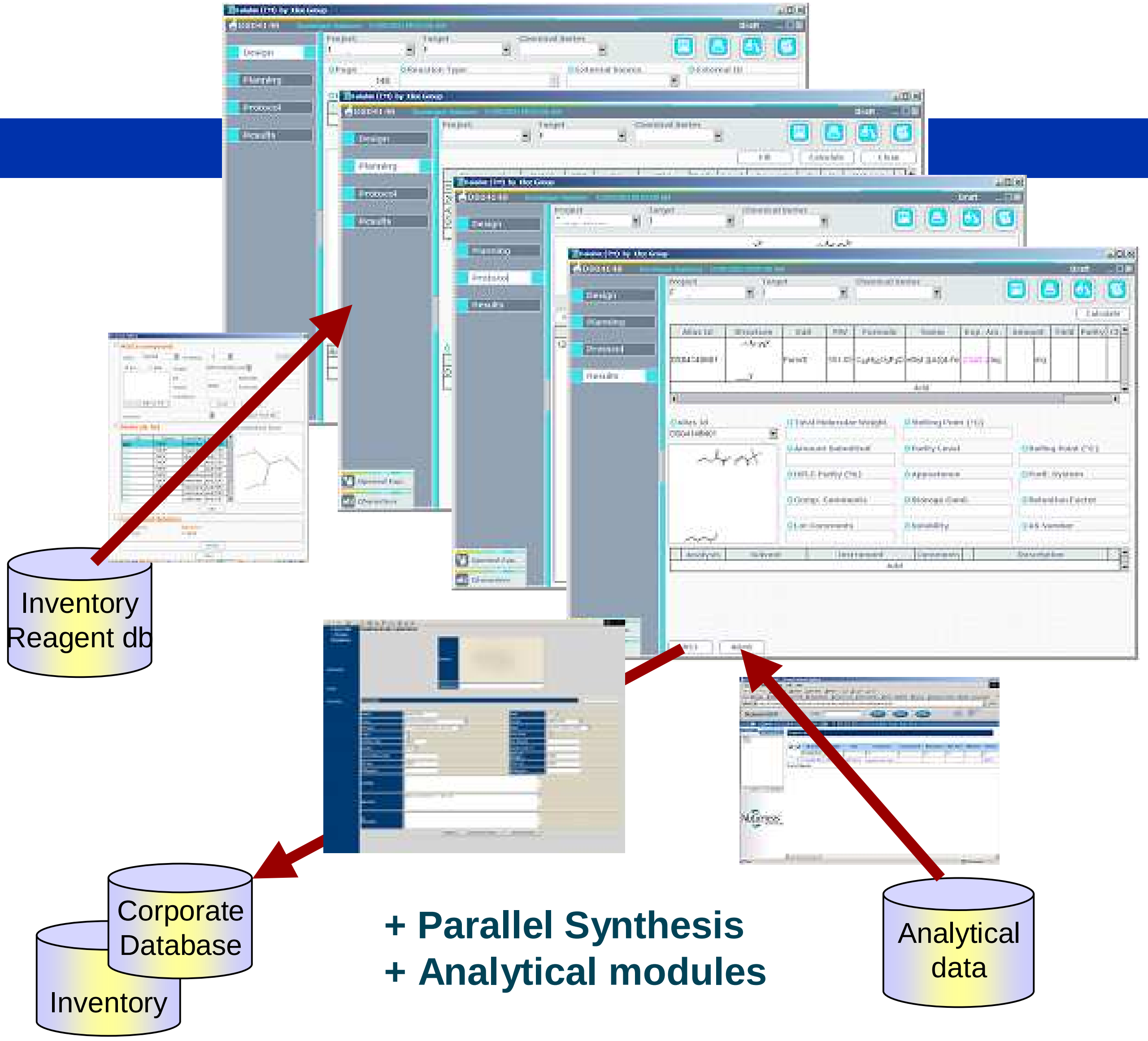
Corporate Database

# Screening data pipeline



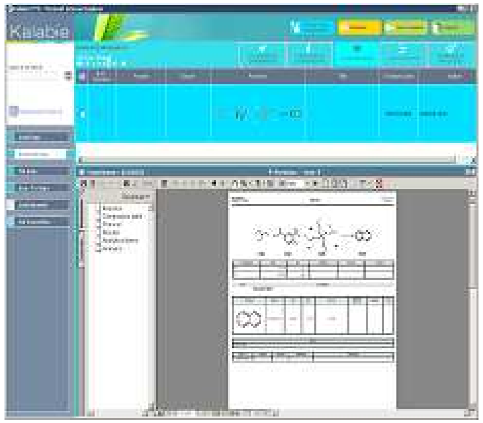
Data is transferred automatically, reducing human mistakes. Screeners can focus on QC.

# Electronic Laboratory Notebook



**+ Parallel Synthesis**
**+ Analytical modules**

**Queries, Reports (e.g. patent writing that automatically re-builds all reaction scheme)**

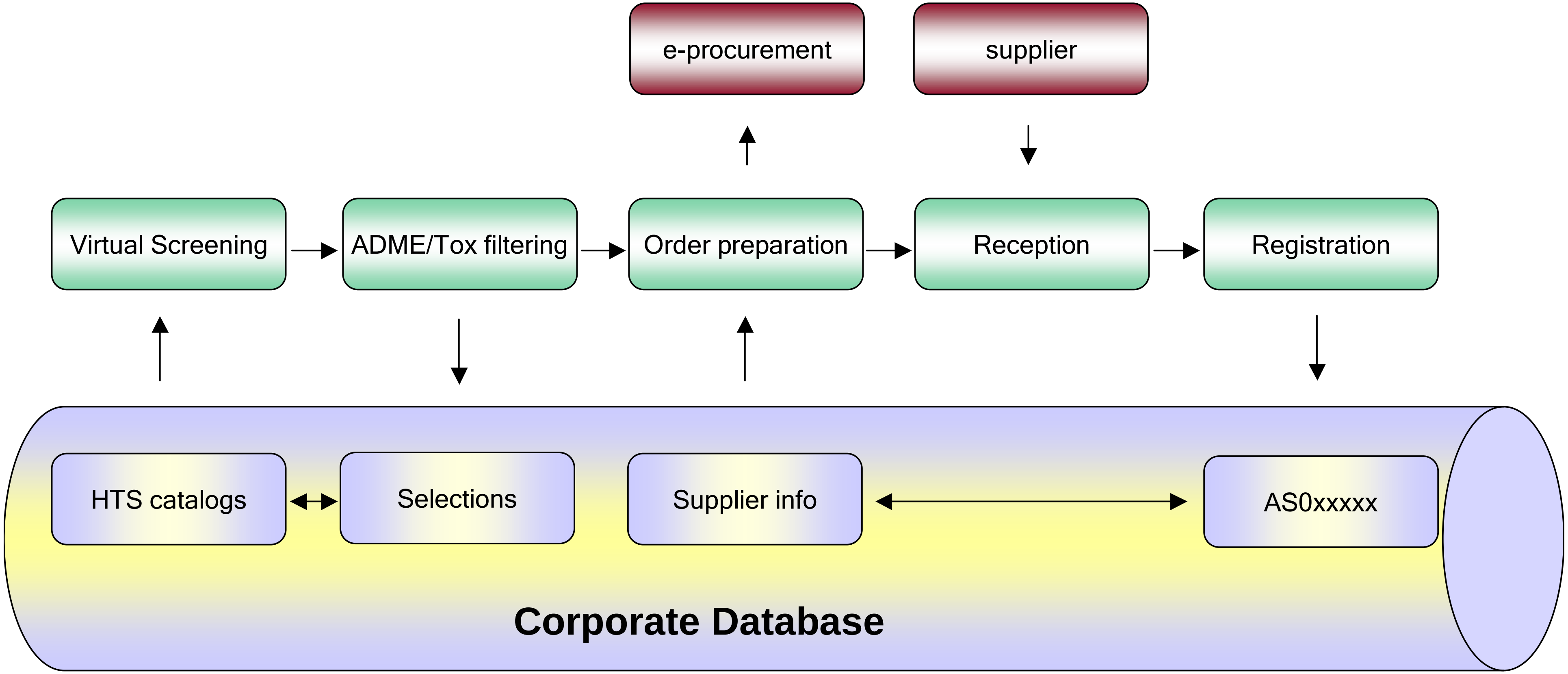


Automatic structure import and calculations, registration in corporate db and inventory ease the chemists’ work

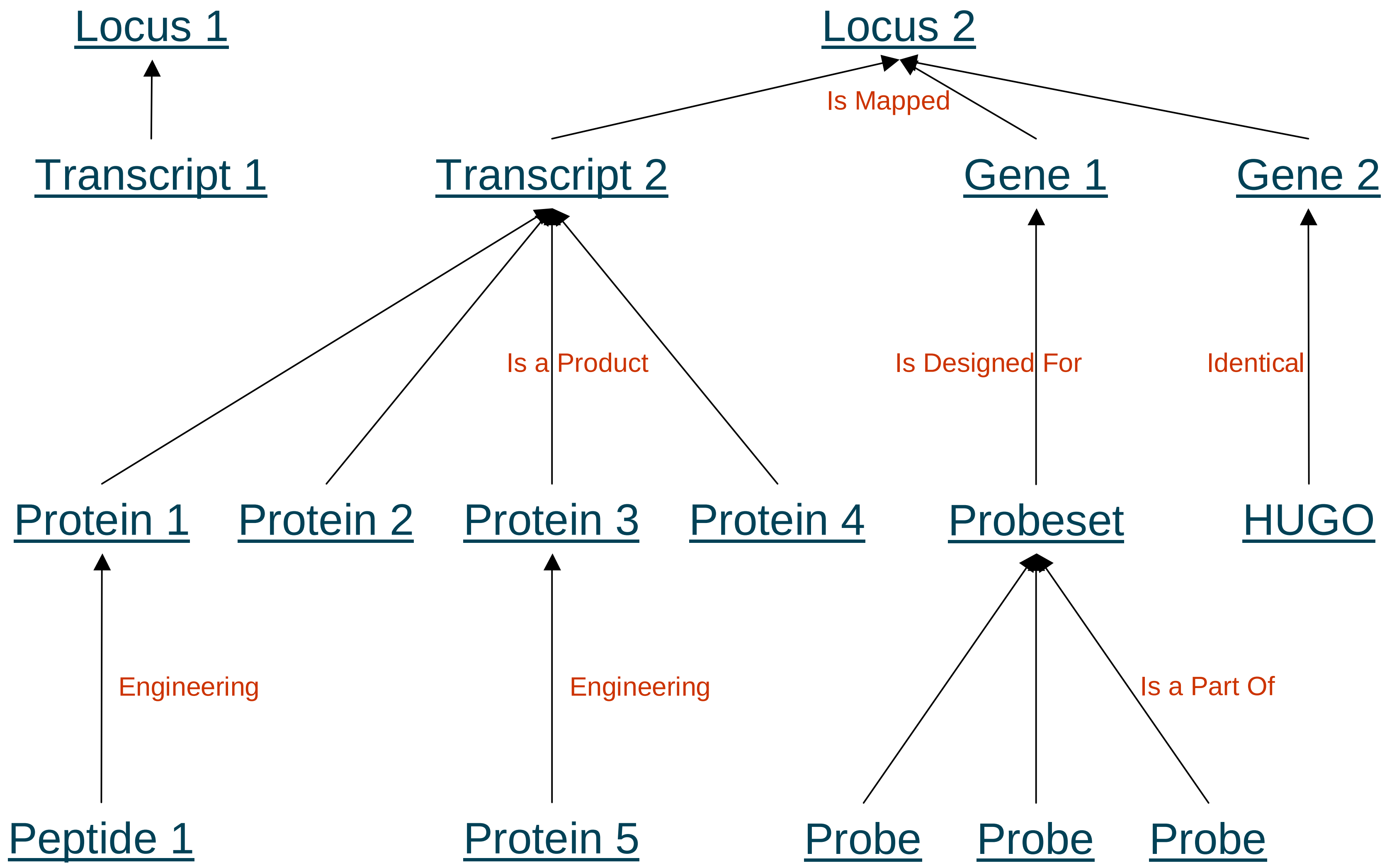
# Workflow Integration

- Project-relevant information must remain in the Corporate DB

- Rule of thumb:
  - Corporate DB is a repository for all entities
  - LIMS processes raw data into information
  - LIMS associates information to entities in the Corporate DB



# Aggregation: Sequences



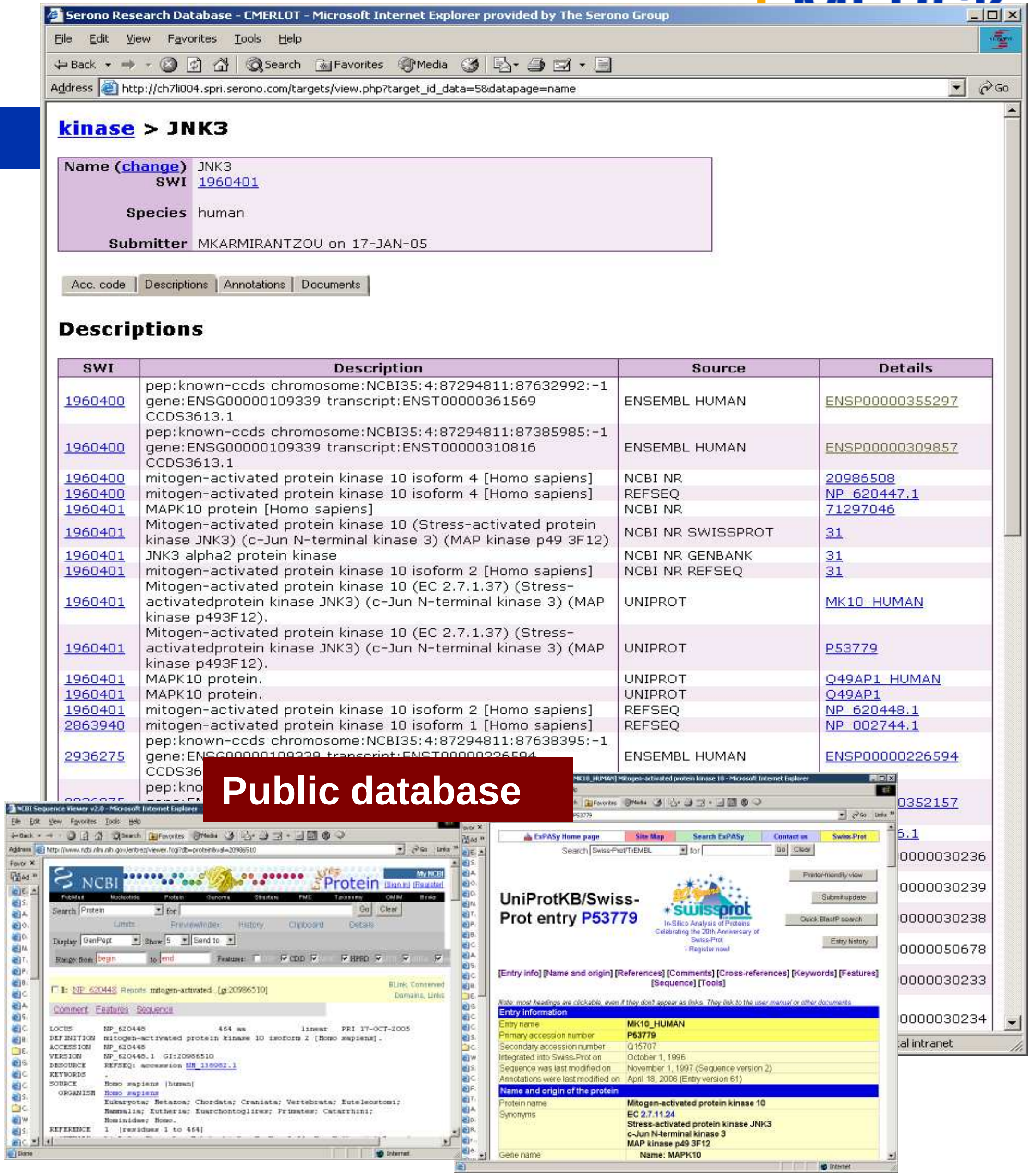
# Application – Protein Database

## Single access to the biological information:

- Description
- Gene expression
- Orthologs
- Disease link
- …

## Grouping under a single protein all the annotations about related genes, transcripts or orthologs

## Access to public and corporate data (e.g. Screening, production)



Serono Research Database - CMERLOT - Microsoft Internet Explorer provided by The Serono Group

Address: http://ch7li004.spri.serono.com/targets/view.php?target_id_data=5&datapage=name

kinase > JNK3

| | |
|---|---|
| Name (change) | JNK3 |
| SWI | 1960401 |
| Species | human |
| Submitter | MKARMIRANTZOU on 17-JAN-05 |

### Descriptions

| SWI | Description | Source | Details |
|---|---|---|---|
| 1960400 | pep:known-ccds chromosome:NCBI35:4:87294811:87632992:-1 gene:ENSG00000109339 transcript:ENST00000361569 CCDS3613.1 | ENSEMBL HUMAN | ENSP00000355297 |
| 1960400 | pep:known-ccds chromosome:NCBI35:4:87294811:87385985:-1 gene:ENSG00000109339 transcript:ENST00000310816 CCDS3613.1 | ENSEMBL HUMAN | ENSP00000309857 |
| 1960400 | mitogen-activated protein kinase 10 isoform 4 [Homo sapiens] | NCBI NR | 20986508 |
| 1960400 | mitogen-activated protein kinase 10 isoform 4 [Homo sapiens] | REFSEQ | NP_620447.1 |
| 1960401 | MAPK10 protein [Homo sapiens] | NCBI NR | 71297046 |
| 1960401 | Mitogen-activated protein kinase 10 (Stress-activated protein kinase JNK3) (c-Jun N-terminal kinase 3) (MAP kinase p49 3F12) | NCBI NR SWISSPROT | 31 |
| 1960401 | JNK3 alpha2 protein kinase | NCBI NR GENBANK | 31 |
| 1960401 | mitogen-activated protein kinase 10 isoform 2 [Homo sapiens] | NCBI NR REFSEQ | 31 |
| 1960401 | Mitogen-activated protein kinase 10 (EC 2.7.1.37) (Stress-activatedprotein kinase JNK3) (c-Jun N-terminal kinase 3) (MAP kinase p493F12). | UNIPROT | MK10_HUMAN |
| 1960401 | Mitogen-activated protein kinase 10 (EC 2.7.1.37) (Stress-activatedprotein kinase JNK3) (c-Jun N-terminal kinase 3) (MAP kinase p493F12). | UNIPROT | P53779 |
| 1960401 | MAPK10 protein. | UNIPROT | Q49AP1_HUMAN |
| 1960401 | MAPK10 protein. | UNIPROT | Q49AP1 |
| 1960401 | mitogen-activated protein kinase 10 isoform 2 [Homo sapiens] | REFSEQ | NP_620448.1 |
| 2863940 | mitogen-activated protein kinase 10 isoform 1 [Homo sapiens] | REFSEQ | NP_002744.1 |
| 2936275 | pep:known-ccds chromosome:NCBI35:4:87294811:87638395:-1 gene:EN CCDS36 | ENSEMBL HUMAN | ENSP00000226594 |
| | pep:kno | | 0352157 |
| | | | 6.1 |
| | | | 0000030236 |
| | | | 0000030239 |
| | | | 0000030238 |
| | | | 0000050678 |
| | | | 0000030233 |
| | | | 0000030234 |

**Public database**

NCBI Sequence Viewer v2.0 - Microsoft Internet Explorer

Address: http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=protein&val=20986510

Search Protein for

Display GenPept Show 5 Send to

Range from begin to end Features: □ ☑ CDD ☑ ☑ HPRD ☑ ☑ ☑

□ 1: NP_620448 Reports mitogen-activated...[gi:20986510]

BLink, Conserved Domains, Links

Comment Features Sequence

```
LOCUS       NP_620448                 464 aa            linear   PRI 17-OCT-2005
DEFINITION  mitogen-activated protein kinase 10 isoform 2 [Homo sapiens].
ACCESSION   NP_620448
VERSION     NP_620448.1  GI:20986510
DBSOURCE    REFSEQ: accession NM_138982.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Catarrhini;
            Hominidae; Homo.
REFERENCE   1  (residues 1 to 464)
```

MK10_HUMAN Mitogen-activated protein kinase 10 - Microsoft Internet Explorer

Search Swiss-Prot/TrEMBL for

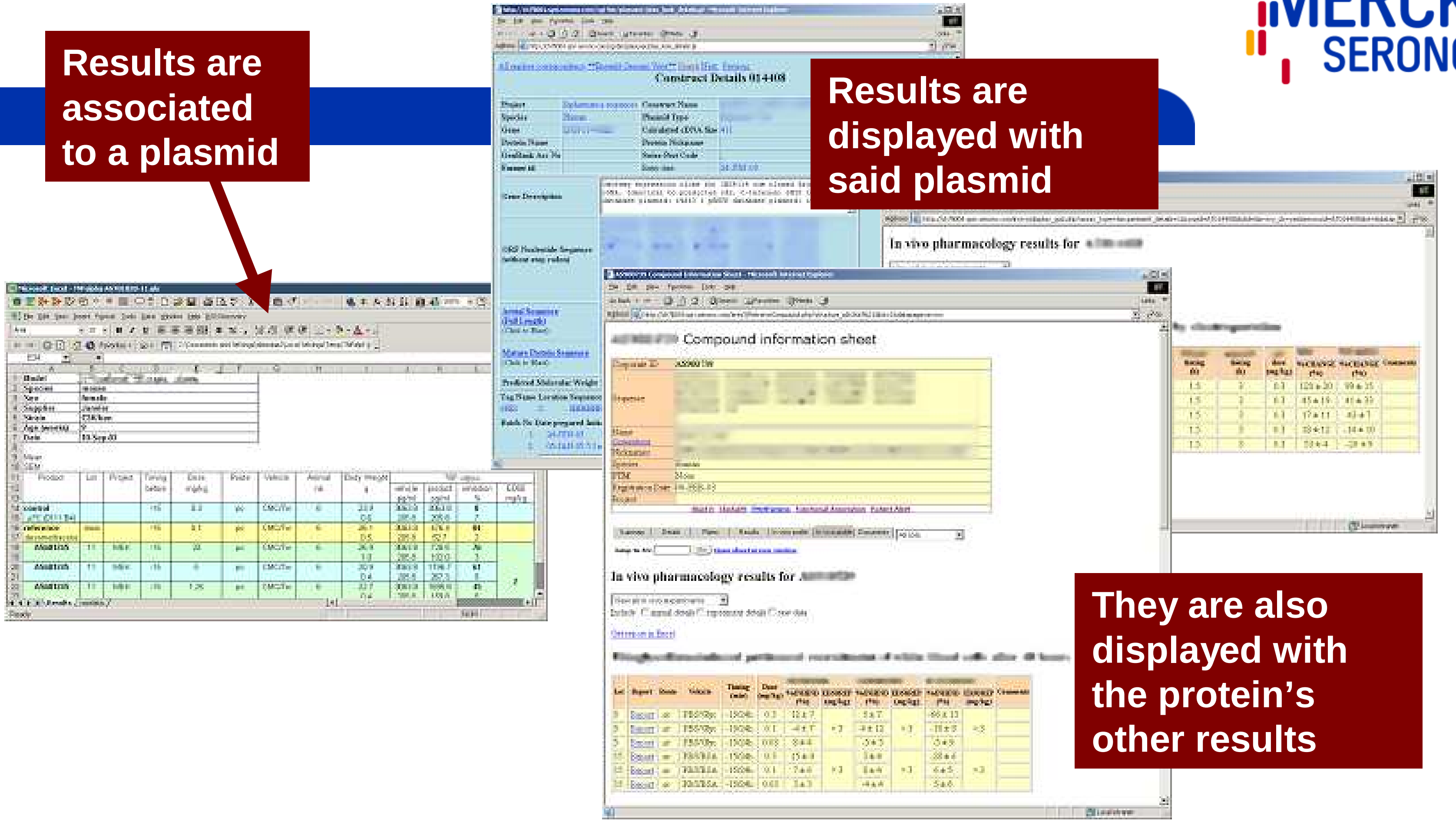
UniProtKB/Swiss-Prot entry P53779

[Entry info] [Name and origin] [References] [Comments] [Cross-references] [Keywords] [Features] [Sequence] [Tools]

Note: most headings are clickable, even if they don't appear as links. They link to the user manual or other documents

| Entry information | |
|---|---|
| Entry name | MK10_HUMAN |
| Primary accession number | P53779 |
| Secondary accession number | Q15707 |
| Integrated into Swiss-Prot on | October 1, 1996 |
| Sequence was last modified on | November 1, 1997 (Sequence version 2) |
| Annotations were last modified on | April 18, 2006 (Entry version 61) |
| **Name and origin of the protein** | |
| Protein name | Mitogen-activated protein kinase 10 |
| Synonyms | EC 2.7.11.24<br>Stress-activated protein kinase JNK3<br>c-Jun N-terminal kinase 3<br>MAP kinase p49 3F12 |
| Gene name | Name: MAPK10 |

# Reporting sequence related data



Construct Details 014408

Compound information sheet

In vivo pharmacology results for

Results are associated to a plasmid

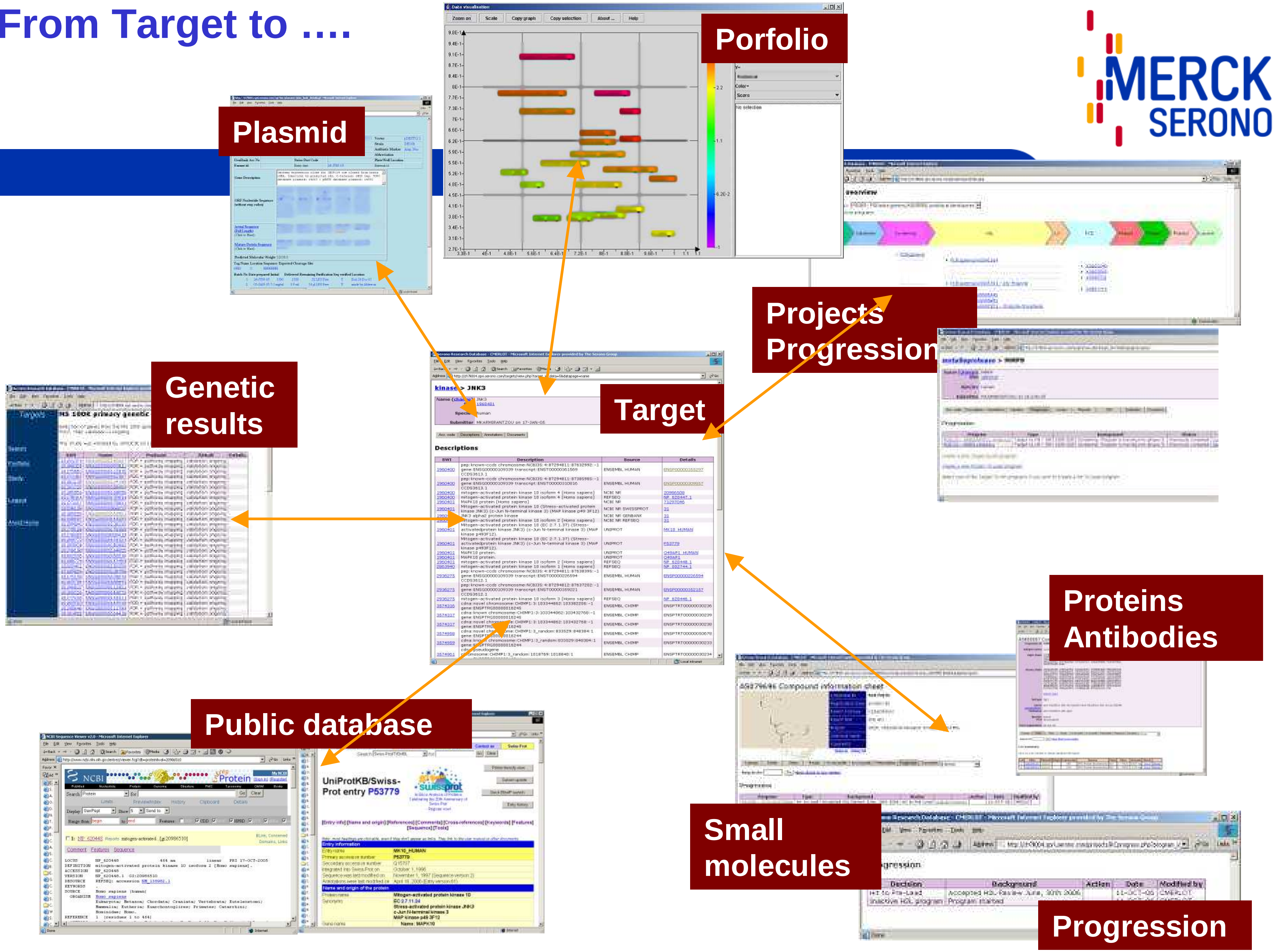
Results are displayed with said plasmid

They are also displayed with the protein’s other results

- Most results are associated to a protein (cell assays, in vivo models) but some are associated to other entities (plasmids,…).
- All must appear in the different contexts.

# From Target to ….

MERCK SERONO



Porfolio

Plasmid

Projects Progression

Genetic results

Target

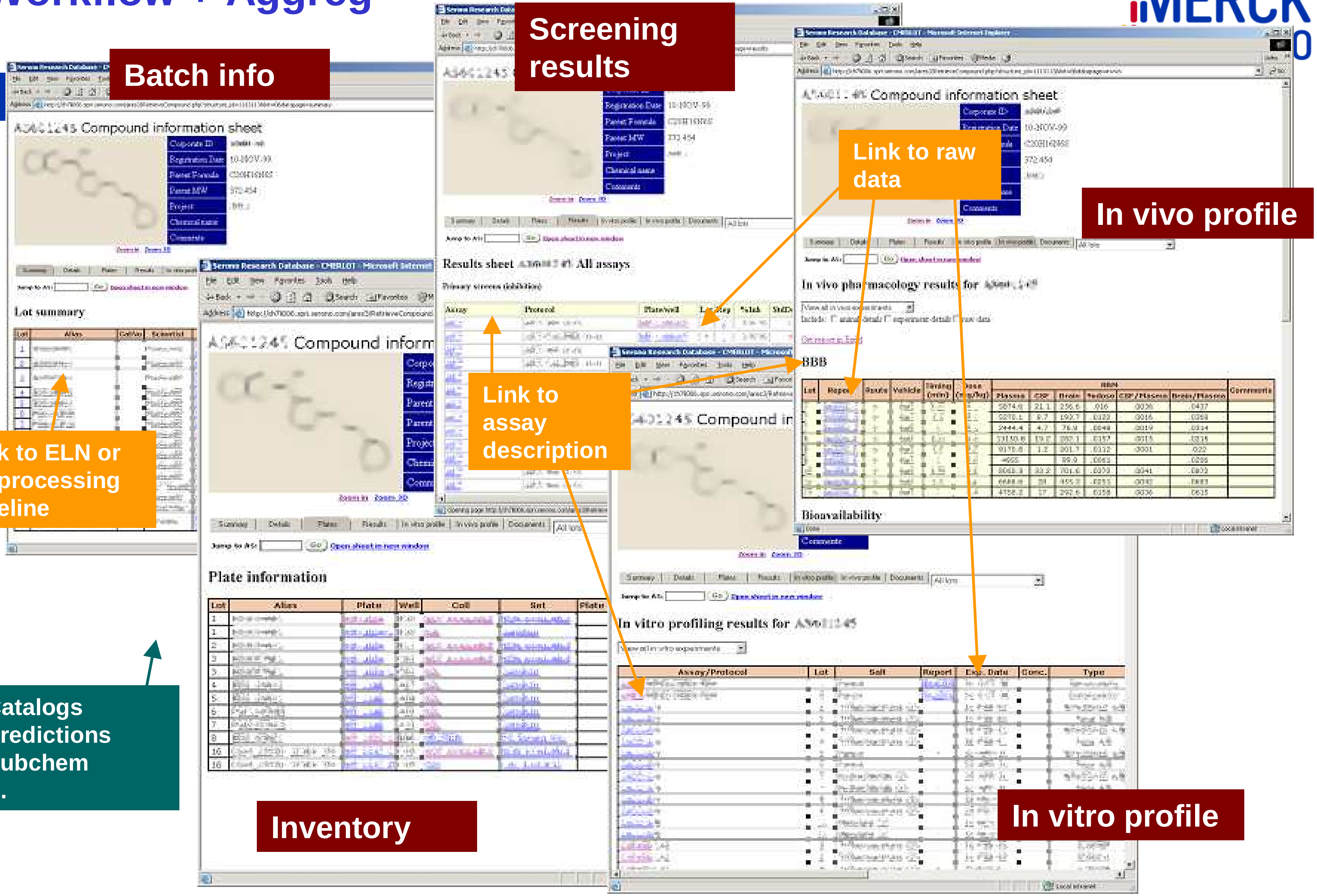
Proteins Antibodies

Public database

Small molecules

Progression

# Molecule Centric Integration: One Compound, All Information Workflow + Aggreg


Compound information sheet
Lot summary
Results sheet All assays
Primary screens (inhibition)
Plate information
In vivo pharmacology results for
BBB
Bioavailability
In vitro profiling results for

Batch info

Screening results

Link to raw data

In vivo profile

Link to ELN or bioprocessing pipeline

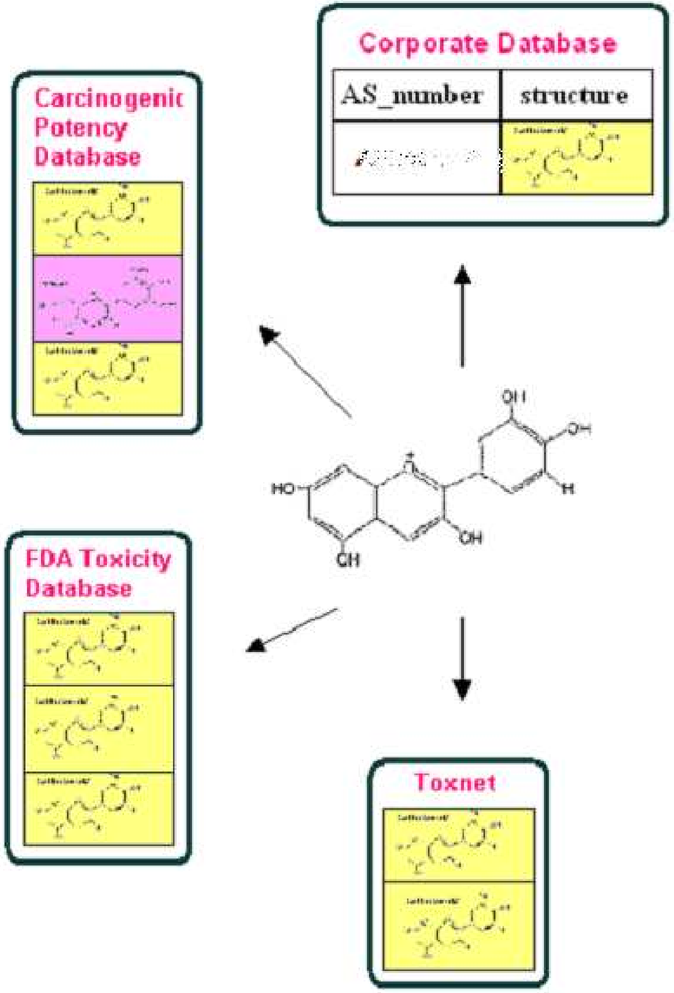
Link to assay description

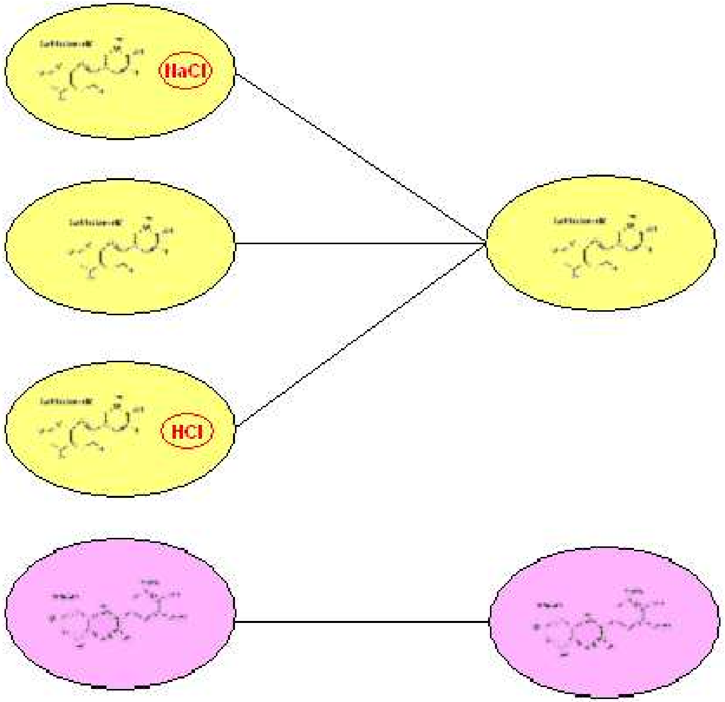
Catalogs
Predictions
Pubchem
…

Inventory

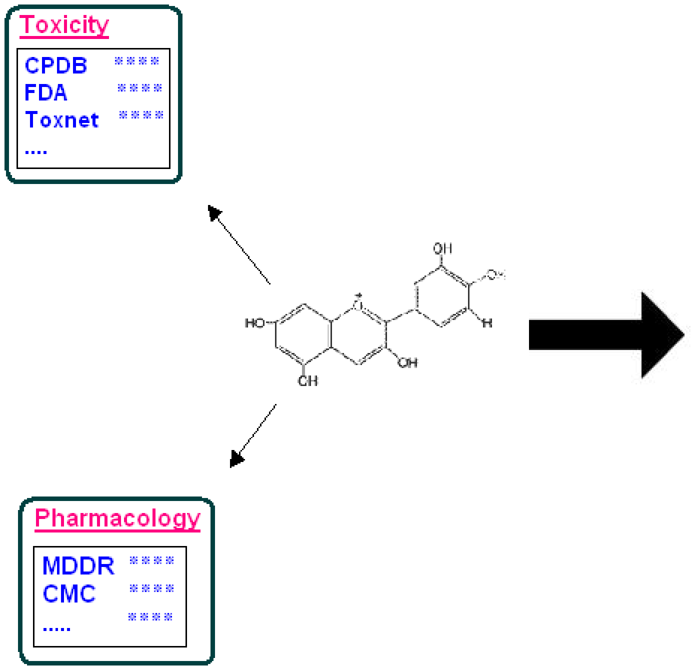
In vitro profile

# Structure Integration





# Result Integration

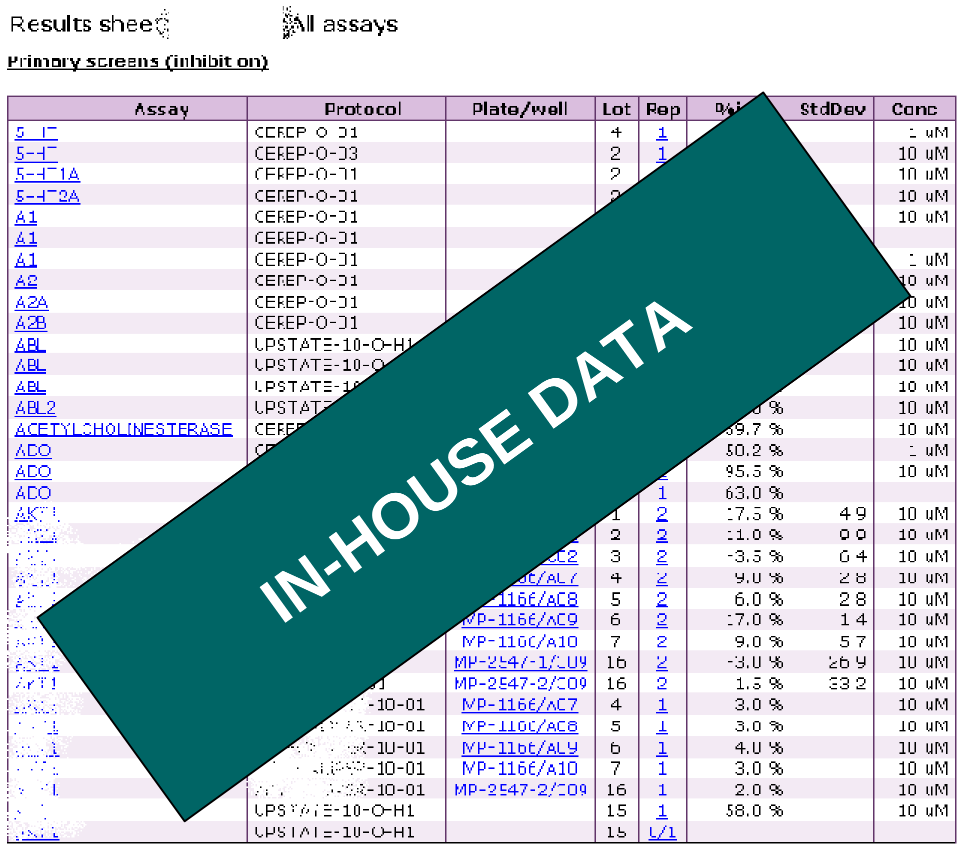


Toxicity

CPDB ****
FDA ****
Toxnet ****
....

Pharmacology

MDDR ****
CMC ****
..... ****



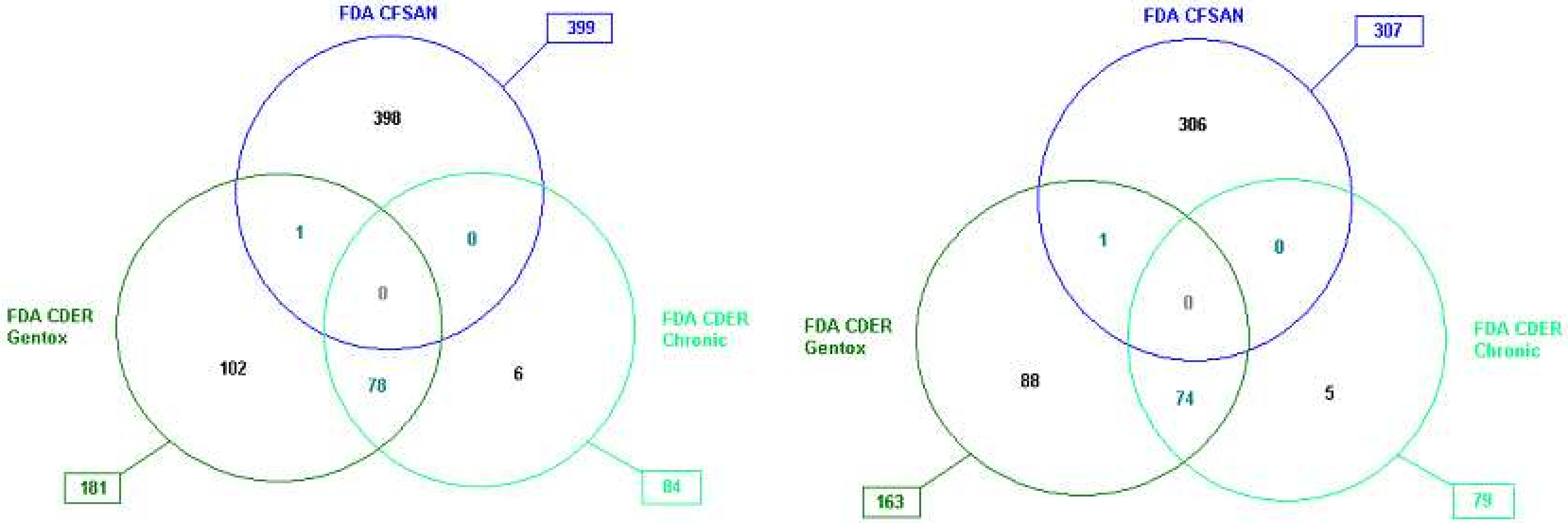
Results sheet  All assays

**Primary screens (inhibition)**

| Assay | Protocol | Plate/well | Lot | Rep | %i | StdDev | Conc |
|---|---|---|---|---|---|---|---|
| 5-HT | CEREP O 01 |  | 4 | 1 | [illegible] | [illegible] | 1 uM |
| 5-HT | CEREP-O-03 |  | 2 | 1 | [illegible] | [illegible] | 10 uM |
| 5-HT1A | CEREP-O-01 |  | 2 | [illegible] | [illegible] | [illegible] | 10 uM |
| 5-HT2A | CEREP-O-01 |  | 2 | [illegible] | [illegible] | [illegible] | 10 uM |
| A1 | CEREP-O-01 |  | [illegible] | [illegible] | [illegible] | [illegible] | 10 uM |
| A1 | CEREP-O-01 |  | [illegible] | [illegible] | [illegible] | [illegible] |  |
| A1 | CEREP-O-01 |  | [illegible] | [illegible] | [illegible] | [illegible] | 1 uM |
| A2 | CEREP-O-01 |  | [illegible] | [illegible] | [illegible] | [illegible] | 10 uM |
| A2A | CEREP-O-01 |  | [illegible] | [illegible] | [illegible] | [illegible] | 10 uM |
| A2B | CEREP-O-01 |  | [illegible] | [illegible] | [illegible] |  | 10 uM |
| ABL | UPSTATE-10-O-H1 | [illegible] | [illegible] | [illegible] | [illegible] |  | 10 uM |
| ABL | UPSTATE-10-O[illegible] | [illegible] | [illegible] | [illegible] | [illegible] |  | 10 uM |
| ABL | UPSTATE-1[illegible] | [illegible] | [illegible] | [illegible] | [illegible] |  | 10 uM |
| ABL2 | UPSTAT[illegible] | [illegible] | [illegible] | [illegible] | [illegible].0 % |  | 10 uM |
| ACETYLCHOLINESTERASE | CERE[illegible] | [illegible] | [illegible] | [illegible] | 59.7 % |  | 10 uM |
| ADO | C[illegible] | [illegible] | [illegible] | [illegible] | 50.2 % |  | 1 uM |
| ADO | [illegible] | [illegible] | [illegible] | [illegible] | 95.5 % |  | 10 uM |
| ADO | [illegible] | [illegible] | [illegible] | 1 | 63.0 % |  |  |
| AKT1 | [illegible] | [illegible] | 1 | 2 | 17.5 % | 4 9 | 10 uM |
| [illegible] | [illegible] | [illegible] | 2 | 2 | 11.0 % | 9 9 | 10 uM |
| [illegible] | [illegible] | [illegible]C2 | 3 | 2 | -3.5 % | 6 4 | 10 uM |
| [illegible] | [illegible] | [illegible]6/AC7 | 4 | 2 | 9.0 % | 2 8 | 10 uM |
| [illegible] | [illegible] | [illegible]-1166/AC8 | 5 | 2 | 6.0 % | 2 8 | 10 uM |
| [illegible] | [illegible] | MP-1166/AC9 | 6 | 2 | 17.0 % | 1 4 | 10 uM |
| [illegible] | [illegible] | MP-1166/A10 | 7 | 2 | 9.0 % | 5 7 | 10 uM |
| AKT1 | [illegible] | MP-2547-1/C09 | 16 | 2 | -3.0 % | 26 9 | 10 uM |
| AKT1 | [illegible]01 | MP-2547-2/C09 | 16 | 2 | 1.5 % | 33 2 | 10 uM |
| [illegible] | [illegible]-10-01 | MP-1166/AC7 | 4 | 1 | 3.0 % |  | 10 uM |
| [illegible] | [illegible]-10-01 | MP-1166/AC8 | 5 | 1 | 3.0 % |  | 10 uM |
| [illegible] | [illegible]-10-01 | MP-1166/AC9 | 6 | 1 | 4.0 % |  | 10 uM |
| [illegible] | [illegible]-10-01 | MP-1166/A10 | 7 | 1 | 3.0 % |  | 10 uM |
| [illegible] | [illegible]-10-01 | MP-2547-2/C09 | 16 | 1 | 2.0 % |  | 10 uM |
| [illegible] | UPSTATE-10-O-H1 |  | 15 | 1 | 58.0 % |  | 10 uM |
| [illegible] | UPSTATE-10-O-H1 |  | 15 | 1/1 |  |  |  |

IN-HOUSE DATA

# FDA Toxicity Database



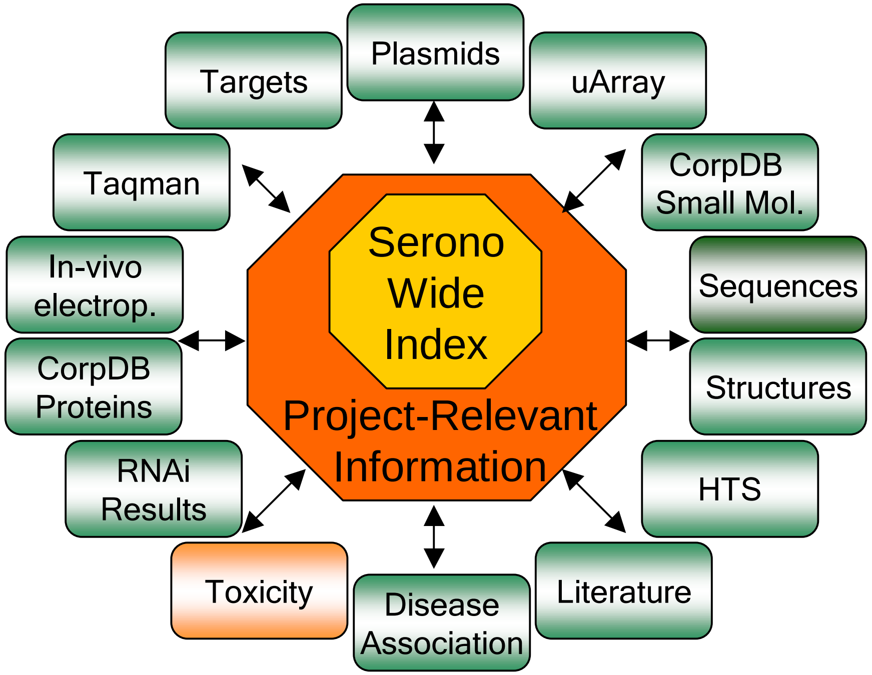
without saltstripping

with saltstripping

Venn diagram of the FDA toxicity databases

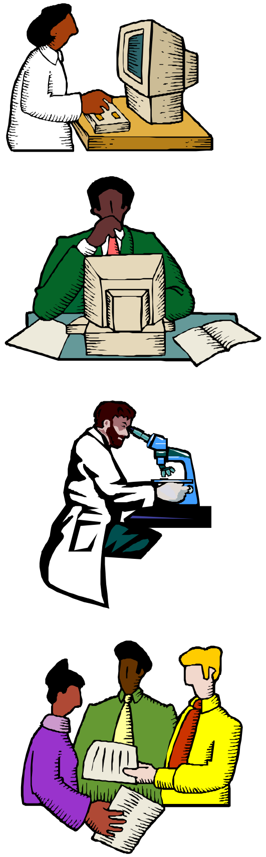
# "Structured" information sources

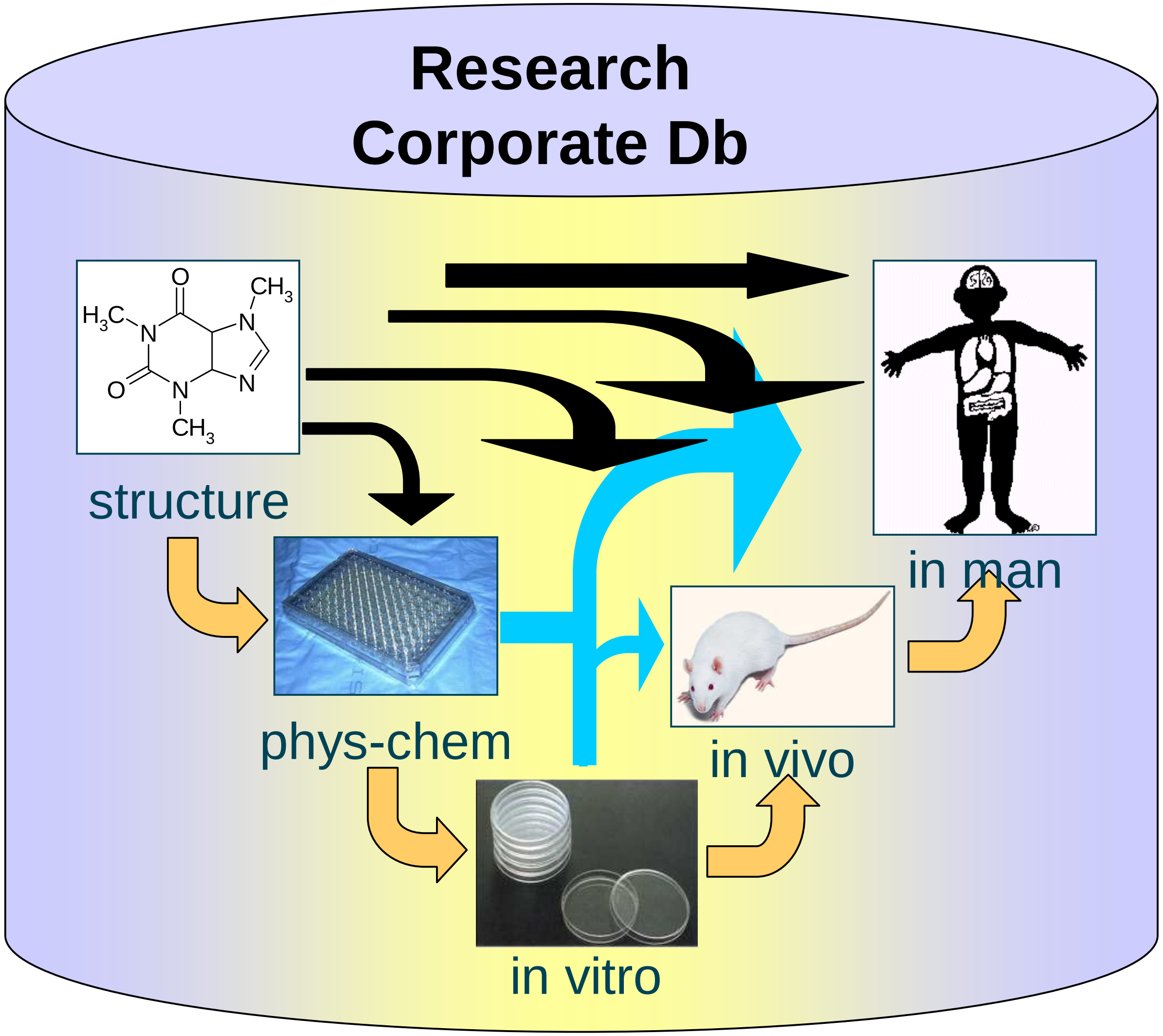
- **Integration requires prior in-depth analysis of contents**
- **Database (relational) <> structured.**
  - Would need to expose model and contents structure
  - No standards nor unique definitions
- **E.g. Toxicity.**
  - Endpoint definition, experimental details.
  - Models: Might predict the same activity but for different reasons.
    - Two different fragments responsible for same activity
    - Activity extrapolated from different experiments



# Needs

- Integrate
- Monitor
- Informed decision-making





# Combining Toxicity Prediction Methods - Theoretical Study

- **Combining predictions increases either sensitivity OR specificity.**

- **Hypothesis:**
  - A substance is toxic if it reaches then binds/reacts to its target. These two requirements are modeled separately. We assume the substance reaches its target.
  - The binding event is due to the arrangement of some atoms in the molecule. The rest of them is not required but might interfere (eg: steric hindrance)

- **Method 1:** results for a single method
  - `false positive rate:      30.97%`
  - `false negative rate:      12.04%`

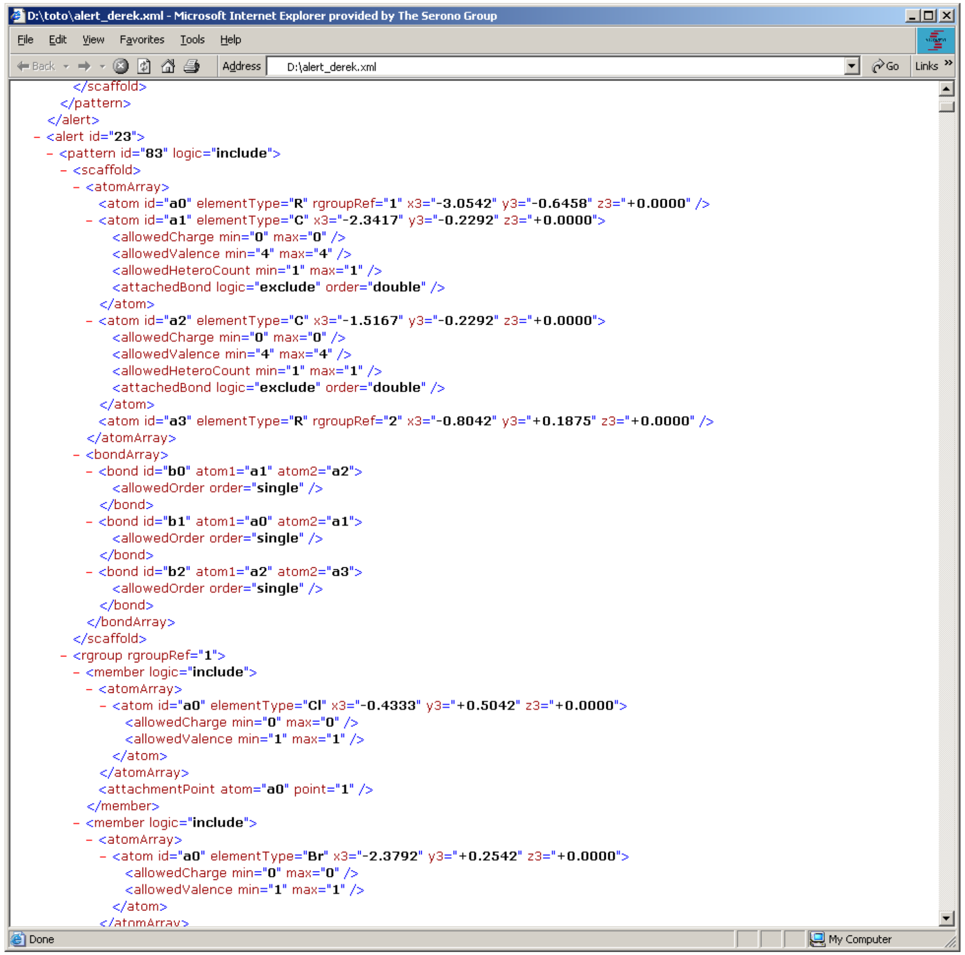
- **Method 2:** predicted positive if A and B are positive with a common descriptor
  - `false positive rate:       0.25%` **=> highest specificity**
  - `false negative rate:      22.41%`

- **Method 3:** predicted positive when both A and B are positive
  - `false positive rate:      12.10%`
  - `false negative rate:      17.87%`

- **Method 4:** predicted positive when either A or B is positive
  - `false positive rate:      48.93%`
  - `false negative rate:       6.52%` **=> highest sensitivity**

Simulation based on evaluating an outcome for 100000 molecules with different theoretical methods. Principle: different methods know some arrangements that cause toxicity, they miss some of them (=> false negatives) and include arrangements that are not a root cause (=> false positives).

# chemXML – sharing alerts



D:\toto\alert_derek.xml - Microsoft Internet Explorer provided by The Serono Group
File Edit View Favorites Tools Help
Back
Address
D:\alert_derek.xml
Go
Links
Done
My Computer

```
         </scaffold>
       </pattern>
     </alert>
   - <alert id="23">
     - <pattern id="83" logic="include">
       - <scaffold>
         - <atomArray>
             <atom id="a0" elementType="R" rgroupRef="1" x3="-3.0542" y3="-0.6458" z3="+0.0000" />
           - <atom id="a1" elementType="C" x3="-2.3417" y3="-0.2292" z3="+0.0000">
               <allowedCharge min="0" max="0" />
               <allowedValence min="4" max="4" />
               <allowedHeteroCount min="1" max="1" />
               <attachedBond logic="exclude" order="double" />
             </atom>
           - <atom id="a2" elementType="C" x3="-1.5167" y3="-0.2292" z3="+0.0000">
               <allowedCharge min="0" max="0" />
               <allowedValence min="4" max="4" />
               <allowedHeteroCount min="1" max="1" />
               <attachedBond logic="exclude" order="double" />
             </atom>
             <atom id="a3" elementType="R" rgroupRef="2" x3="-0.8042" y3="+0.1875" z3="+0.0000" />
           </atomArray>
         - <bondArray>
           - <bond id="b0" atom1="a1" atom2="a2">
               <allowedOrder order="single" />
             </bond>
           - <bond id="b1" atom1="a0" atom2="a1">
               <allowedOrder order="single" />
             </bond>
           - <bond id="b2" atom1="a2" atom2="a3">
               <allowedOrder order="single" />
             </bond>
           </bondArray>
         </scaffold>
       - <rgroup rgroupRef="1">
         - <member logic="include">
           - <atomArray>
             - <atom id="a0" elementType="Cl" x3="-0.4333" y3="+0.5042" z3="+0.0000">
                 <allowedCharge min="0" max="0" />
                 <allowedValence min="1" max="1" />
               </atom>
             </atomArray>
             <attachmentPoint atom="a0" point="1" />
           </member>
         - <member logic="include">
           - <atomArray>
             - <atom id="a0" elementType="Br" x3="-2.3792" y3="+0.2542" z3="+0.0000">
                 <allowedCharge min="0" max="0" />
                 <allowedValence min="1" max="1" />
               </atom>
             </atomArray>
```

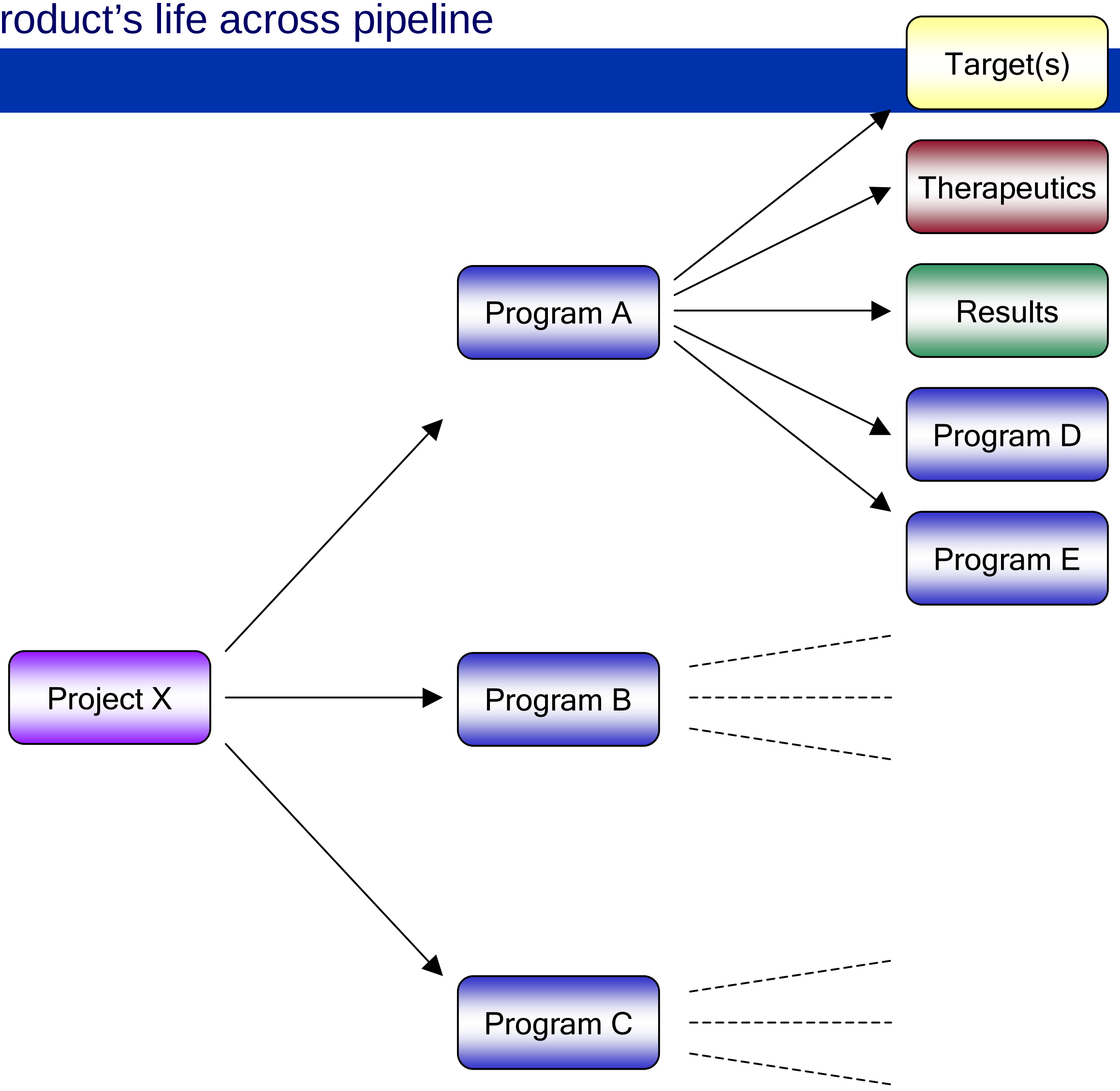
- Extends CML to substructure queries
  http://cml.sourceforge.net
- Handles almost all substructure query features described in MOL files, Derek and SMARTS
- Text-based format for sharing easily query definition
- Most of the validation is done by the XML parser (eg: Xerces)

# Concluding remarks

- **Unique Id for entities**
  - Advantages: Fast and reliable. "Easy" integration for in-house information
  - Drawbacks:
    - Indefinitely scalable?:
      - chemical space estimated to be 10^100 individual structures. Not incl mixtures.
      - Nessea: Developed algorithm to avoid the enumeration of Combichem Libraries
- **Production workflow tracking**
- **Project relevant information in corporate database**
- **Horizontal and vertical navigation**
- **Main trends:**
  - ELNs for know-how capitalization.
  - Information scoring/mapping
  - Extend scope. Cover whole life of entities.
  - Exploit data from outside
- **Main issue to allow integration and automated of parsing/interpretation:**
  - Lack of unique definitions/standards to depict entities, relations, contents

# Backup slide: Project Management

- Integrating project-related data
- Tracking product’s life across pipeline



Target(s)

Therapeutics

Results

Program D

Program E

Program A

Project X

Program B

Program C