



Workflow Automation and Data Management

Markus Kroemer

PSDI Conference 2013



Why Automation?

- Increase productivity and efficiency
- Increase quality of products and services

Potential Risks of Automation

- Drop of skills
- Drop of flexibility
- Drop of data quality

Data Processing Pipeline

Project Management Layer

Pipedream

autoPROC

autoBUSTER

XDS

Aimless

Buster/TNT

Power of Semi-Automation

- Halt the pipeline at decisive points
- Enable the user to analyze the data
- Let the expert make the decision


Standardization Enables

- To capture results automatically
- To search and secure information easily
- To manage samples efficiently

Why Controlled Vocabulary

- Crystallization conditions in PDB
 - >25 different annotations of ammonium sulfate
 - How do you do data mining on conditions?
- The solution is controlled vocabulary

Recipe Database



PBS Buffer (1X)
137mM NaCl + 2.7mM KCl + 10mM Na2HPO4·2H2O
+ 2mM KH2PO4
S1458 admin 19/02/2013

Show All Records | *Print Label:* Long Short Cap Tube | Print Report

Save Solution | New Solution | Copy Solution | Delete Solution

ProjectX_ITC_Buffer

ID S1270

Details

Name Buffer

Process Biophysics

Usage ITC

Type Buffer

Project ProjectX

Filtered yes no

Storage +25C

Creator izaacau1

Date 03/02/2012

Status active inactive

Stock Solution

Stock yes no

Add Ingredient **Add Instruction** **Volume** 1000 mL

Type	Final Conc	Ingredient	Toxic <input type="checkbox"/>	Recipe
Stock	20 mM	Bis-Tris propane (1M pH8.5)		20 mL
Stock	100 mM	Na chloride (5M)		20 mL
Organic	2 mM	TCEP		0.5733 g
Detergent	0.1 %w/v	CHAPS		1 g
		Adjust pH to 8.5		

Comment on Label

Comment

Crosslinking of databases

- Enabled by careful curation of unique IDs associated with samples and experiments
 - Uniprot/Gene ID
 - Protein sequence
 - Protein Batch ID
 - Compound ID
 - Crystal ID

End-point Databases

- Most of the database we interact with are “end-point” databases
- Need for quality control and data curation

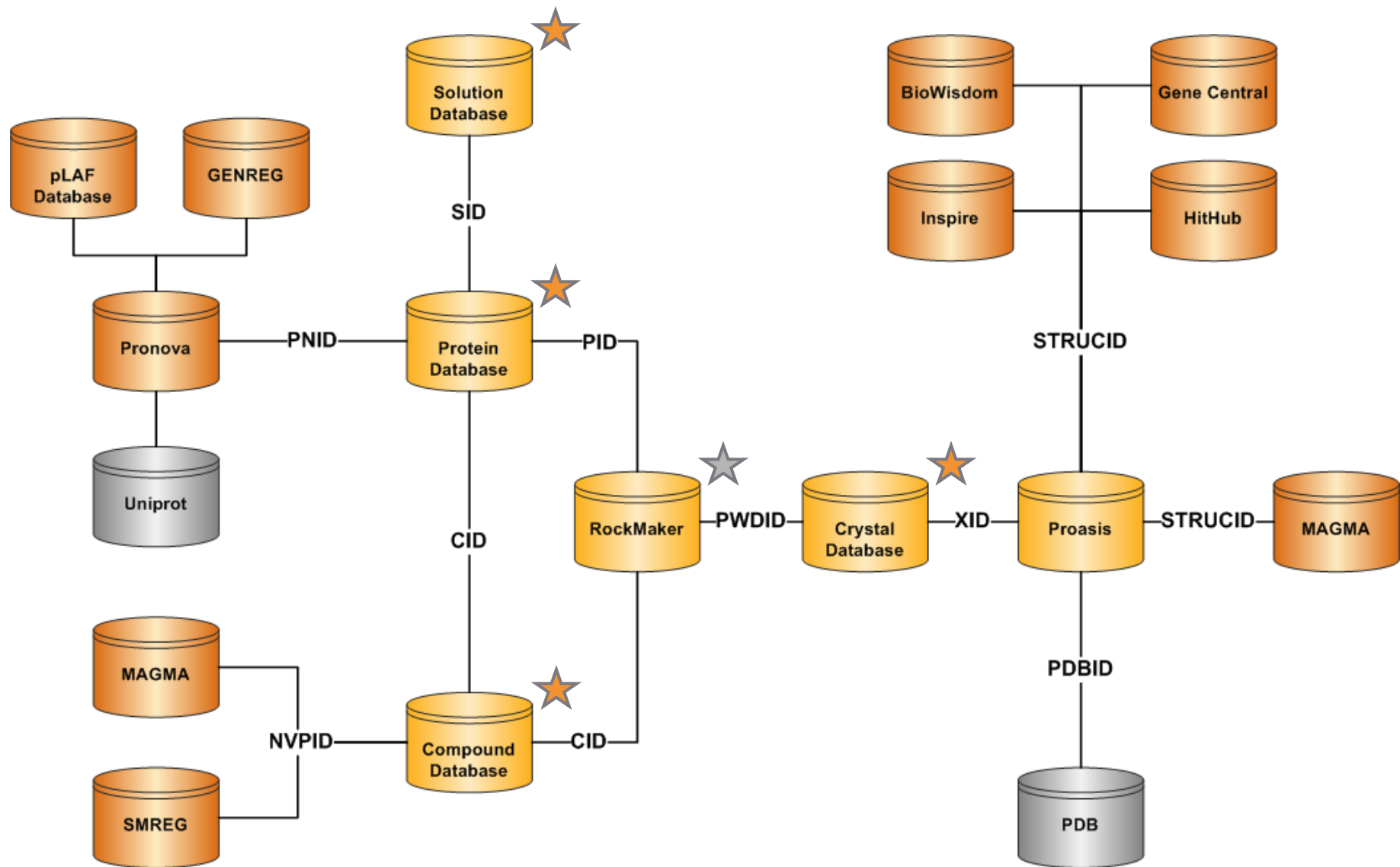
How you can make a difference

- Do not organize or annotate data in a way that helps just you just for the time being
- Think how others will make use of the information
- Think in a broader context not just in the context of the individual sample

In-Process Databases

- Providing benefit to depositors
 - Scientists want „in-process” data capturing tools which help them to be more efficient and which are used to seamlessly upload data into end-point databases

Data Management Concept



Compound Database

Start | Compound Registration | Compounds | Solution Registration | Solutions | Find & Print | Database Menu

list | find by tube | new registration | find | show all | print report | register compounds | list compounds | delete registration | [admin] re-login | help

Compound Registration Details

Rack Scanner (for Compounds ordered from Novartis Archive) [admin]

Scanner

Register Compounds

RegID: 11877 Registered: yes

Operator: waegece1 Reg. Date: 19/02/2013

Temp.[°C]: 4 Rationale: Biophysics Project: gt

Scanner Reads: 70 No Reads: 0 No Tubes: 26

Tube No	NovartisID	Pos.	Molecular Weight eff. [g/mol]	Mass in Tube [mg]	Temp. [°C]	Rationale	Project	Order Name
0124973035	2-NX-1	A 1	331.74	1.60	4	Biophysics	gt	Celine Be
0125053139	5-AO-1	A 2	362.4224	0.70	4	Biophysics	gt	Celine Be
0125053115	06E-NX-1	A 3	320.22	0.70	4			
0124973091	158-AA-1	A 4	354.2803	1.30	4			
0125053142	83-NX-1	A 5	259.36	1.80	4			
0125053141	95-NX-1	A 6	225.25	0.70	4			
0056122106	524-NX-1	A 7	334.33	1.60	4			
0078691760	28-NX-1	A 8	344.38	1.70	4			
0124973157	67-NX-1	A 9	285.78	1.20	4			
0124972487	80-NX-1	A 10	277.72	1.00	4			
0056122157	933-NX-1	A 11	295.18	0.50	4			
0124972983	12-NX-1	A 12	191.23	1.10	4			
0124973139	55A-AA-1	B 1	356.3017	1.40	4			
0124973442	658-NX-1	B 2	268.36	0.90	4			
0124973380	428-NX-1	B 3	350.25	0.80	4			
0124973418	723-NX-1	B 4	322.15	0.60	4			
0124973452	935-NX-1	B 5	297.32	0.90	4			
0124973429	624-NX-1	B 6	320.31	1.10	4			

Count: 69

Start | Compound Registration | Compounds | Solution Registration | Solutions | Find & Print | Database Menu

list | find by tube | new registration | find | show all | print report | register solutions | list solutions | delete registration | [admin] re-login | help

Solution Registration Details

3 [904/904]

Register Solutions

RegID: 11260 Registered: yes

Operator: waegece1 Reg. Date: 19/02/2013

Conc.[mM]: 50 Solvent: 90% dDMSO Temp.[°C]: All

Novartis ID	Operator	Molecular Weight eff. [g/mol]	Mass in Tube [mg]	Conc. [mM]	Volume [uL]	Solvent	Temp. [°C]
2-NX-1	waegece1	331.74	1.60	50	96.5	90% dDMSO	4
6-AO-1	waegece1	362.42	0.70	50	38.6	90% dDMSO	4
06E-NX-1	waegece1	320.22	0.70	50	43.7	90% dDMSO	4
158-AA-1	waegece1	354.28	1.30	50	73.4	90% dDMSO	4
283-NX-1	waegece1	259.36	1.80	50	138.8	90% dDMSO	4
95-NX-1	waegece1	225.25	0.70	50	62.2	90% dDMSO	4
524-NX-1	waegece1	334.33	1.60	50	95.7	90% dDMSO	4
28-NX-1	waegece1	344.38	1.70	50	98.7	90% dDMSO	4
67-NX-1	waegece1	285.78	1.20	50	84.0	90% dDMSO	4
80-NX-1	waegece1	277.72	1.00	50	72.0	90% dDMSO	4
933-NX-1	waegece1	295.18	0.50	50	33.9	90% dDMSO	4
12-NX-1	waegece1	191.23	1.10	50	115.0	90% dDMSO	4
55A-AA-1	waegece1	356.3	1.40	50	78.6	90% dDMSO	4
658-NX-1	waegece1	268.36	0.90	50	67.1	90% dDMSO	4
428-NX-1	waegece1	350.25	0.80	50	45.7	90% dDMSO	4
723-NX-1	waegece1	322.15	0.60	50	37.2	90% dDMSO	4
935-NX-1	waegece1	297.32	0.90	50	60.5	90% dDMSO	4
624-NX-1	waegece1	320.31	1.10	50	68.7	90% dDMSO	4

Count: 69 Max Vol: 195.1


fill out else error

Crystal Database

close
find plate
help

Selected Working Crystal
 Crystal ID Select Crystal

Selected Drop
 Rockmaker ID **RM28959** Well **C3** Drop **1**

Selected Puck
S10


occupied
 selected

Slot	XID	Compound
1	X985	NVP-AWP796-NX-1
2	X986	NVP-AWP796-NX-1

create working crystal
create final crystal
close

Plate View
List View

	1	2	3	4	5	6
A	1 A 1	1 A 2	1 A 3	1 A 4	1 A 5	1 A 6
B	1 B 1	1 B 2	1 B 3	1 B 4	1 B 5	1 B 6
C	1 C 1	1 C 2	1 C 3	1 C 4	1 C 5	1 C 6
D	1 D 1	1 D 2	1 D 3	1 D 4	1 D 5	1 D 6

Soaking
 Compound **NVP-AWP796-NX-1** Conc. **5 mM** Time **1 min**
 Comment
 Soaking comment
 Liquor well other Soak Buffer Drop original new drop Vol.

Cryo Protection
 Protection **glycerol** Conc. **10%** Time **10 sec**
 Comment
 Cryo comment
 Liquor well other Cryo Buffer Drop original new drop Vol.

Data Collection
 Comment
 Collect if better than 2.5Å
 LabHead **kroemma1** SpaceGroup UnitCell

Plate Storage
 Plate ID Well Drop Comment

Rockmaker	Exp. Notes	Protein Formulation	Protein Ingredients	Protein Notes	Well Formulation
RockmakerID RM28959	Notebook notebook	Well C3 - Drop 1 - Volume 1.0uL p + 1.0uL w			Protein Lot PN27692
Date 14/11/2012	Temp [°C] 16	Type Protein	Name ProteinX(110-325)_Y210A	ID P454	Lot 10.0 mg/ml
User Berthold	Seeding <input checked="" type="checkbox"/>	Ligand	CompoundY	C3111	0.6999 mM
Experiment Name Screen M&M compoundY	Plate Type 24 VDX Large	Additive	AdditiveZ		

100 Browse

Interface to Data Collection Service

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	DewarName	PositionInDewar	PuckType	PuckName	PositionInPuck	OperatorID	LabHeadID	ProjectName	ProteinName	CompoundName	CrystalID	Priority	Backup	CrystalGroup	PuckGroup	OperatorComment	Tested? *	DataSet? *	Resolution	Results
1																				
62	Dewar 03	4	MDL	X01	1	waegece1	scheuc1	PE25	CE15	C10327	X801	1		26						
63	Dewar 03	4	MDL	X01	2	waegece1	scheuc1	PE25	CE15	C10327	X802	b	X801	26		please test if x801 is not better				
64	Dewar 03	4	MDL	X01	3	waegece1	scheuc1	PE25	CE15	C10326	X806	1		25						
65	Dewar 03	4	MDL	X01	4	waegece1	scheuc1	PE25	CE15	C10326	X807	b	X806	25		please test if x806 is not better				
66	Dewar 03	4	MDL	X01	5	waegece1	scheuc1	PE25	CE15	C10286	X810	1		27						
67	Dewar 03	4	MDL	X01	6	waegece1	scheuc1	PE25	CE15	C10286	X838	b	X810	27		please test if x810 is not better				
68	Dewar 03	4	MDL	X01	7	waegece1	scheuc1	PE25	CE15	C10287	X839	2		28						
69	Dewar 03	4	MDL	X01	8	waegece1	scheuc1	PE25	CE15	C10287	X840	b	X839	28		please test if x839 is not better				
70	Dewar 03	5	MDL	D10	1	waegece1	kallenjo	PE5	CE4	C10322	X814	1		10						
71	Dewar 03	5	MDL	D10	2	waegece1	kallenjo	PE5	CE4	C10322	X815	b	X814	10		please test if x814 is not better				
72	Dewar 03	5	MDL	D10	3	waegece1	kallenjo	PE5	CE4	C10322	X816	1		10						
73	Dewar 03	5	MDL	D10	4	waegece1	kallenjo	PE5	CE4	C10322	X817	b	X816	10		please test if x816 is not better				
74	Dewar 03	5	MDL	D10	5	waegece1	kallenjo	PE5	CE4	C10334	X818	1		7						
75	Dewar 03	5	MDL	D10	6	waegece1	kallenjo	PE5	CE4	C10334	X819	b	X818	7		please test if x818 is not better				
76	Dewar 03	5	MDL	D10	7	waegece1	kallenjo	PE5	CE4	C10299	X820	1		8						
77	Dewar 03	5	MDL	D10	8	waegece1	kallenjo	PE5	CE4	C10299	X821	b	X820	8		please test if x820 is not better				
78	Dewar 03	5	MDL	D10	9	waegece1	kallenjo	PE5	CE4	C10301	X822	1		9						
79	Dewar 03	5	MDL	D10	10	waegece1	kallenjo	PE5	CE4	C10301	X823	b	X822	9		please test if x822 is not better				
80	Dewar 04	1	MDL	B18	1	zinkfl1	zouch3	PE89	CE12		X830	1		22		salt?				
81	Dewar 04	1	MDL	B18	2	zinkfl1	zouch3	PE89	CE12		X831	1		22		salt?				
82	Dewar 04	1	MDL	B18	3	zinkfl1	zouch3	PE89	CE12		X832	b	X831	22		salt?				
83	Dewar 04	1	MDL	B18	4	zinkfl1	zouch3	PE89	CE12		X833	1		22		salt?				
84	Dewar 04	1	MDL	B18	5	zinkfl1	zouch3	PE89	CE12		X834	b	X834	22		salt?				
85	Dewar 04	1	MDL	B18	6	zinkfl1	zouch3	PE89	CE12		X835	b		22		salt?				
86	Dewar 04	1	MDL	B18	7	zinkfl1	osterni1	PE75	CE7	C9936	X850	1		13		salt?				
87	Dewar 04	1	MDL	B18	8	zinkfl1	osterni1	PE75	CE7	C9936	X851	b	X850	13		salt?				
88	Dewar 04	1	MDL	B18	9	zinkfl1	osterni1	PE20	CE8	C8973	X852	1		15	B18, C12					
89	Dewar 04	1	MDL	B18	10	zinkfl1	osterni1	PE20	CE8	C8973	X853	b	X852	15	B18, C12					
90	Dewar 04	3	MDL	C12	1	zinkfl1	osterni1	PE20	CE8	C8973	X855	1		15	B18, C12					
91	Dewar 04	3	MDL	C12	2	zinkfl1	osterni1	PE20	CE8	C8973	X856	b	X855	15	B18, C12					

Conclusion #1

- Fully automated software pipelines are powerful expert systems which however require resource intense development
- Semi-automated pipelines can be a good compromise taking away input/output formatting from the user but keeping decision making under control of the expert

Conclusion #2

- Standardization enables automation and process optimization

Conclusion #3

- Controlled vocabulary and curated unique dataset identifiers are key for data quality, database cross-linking and data mining

Conclusion #4

- Take the extra minute to annotate your data and results carefully – it's for “eternity”

Conclusion #5

- High quality databases can be achieved by
 - Tailor-made “in-process” data capturing tools
 - Constructive feedback to depositors
 - Dedicated database curators

Thank you for your attention