Workflow Automation and Data Management

Markus Kroemer

PSDI Conference 2013

NOVARTIS

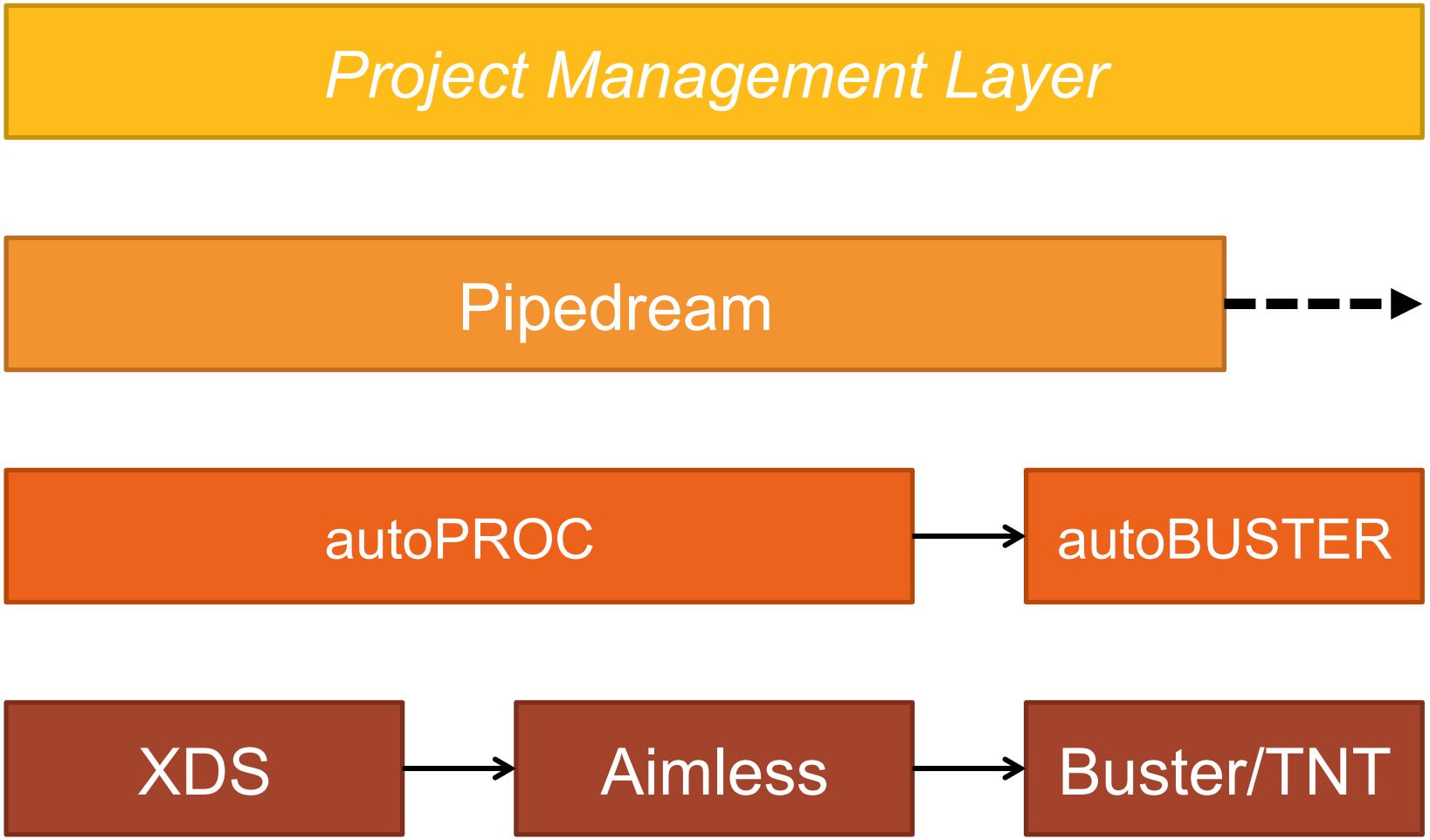
Why Automation?

- 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



NOVARTIS

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	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

NOVARTIS

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

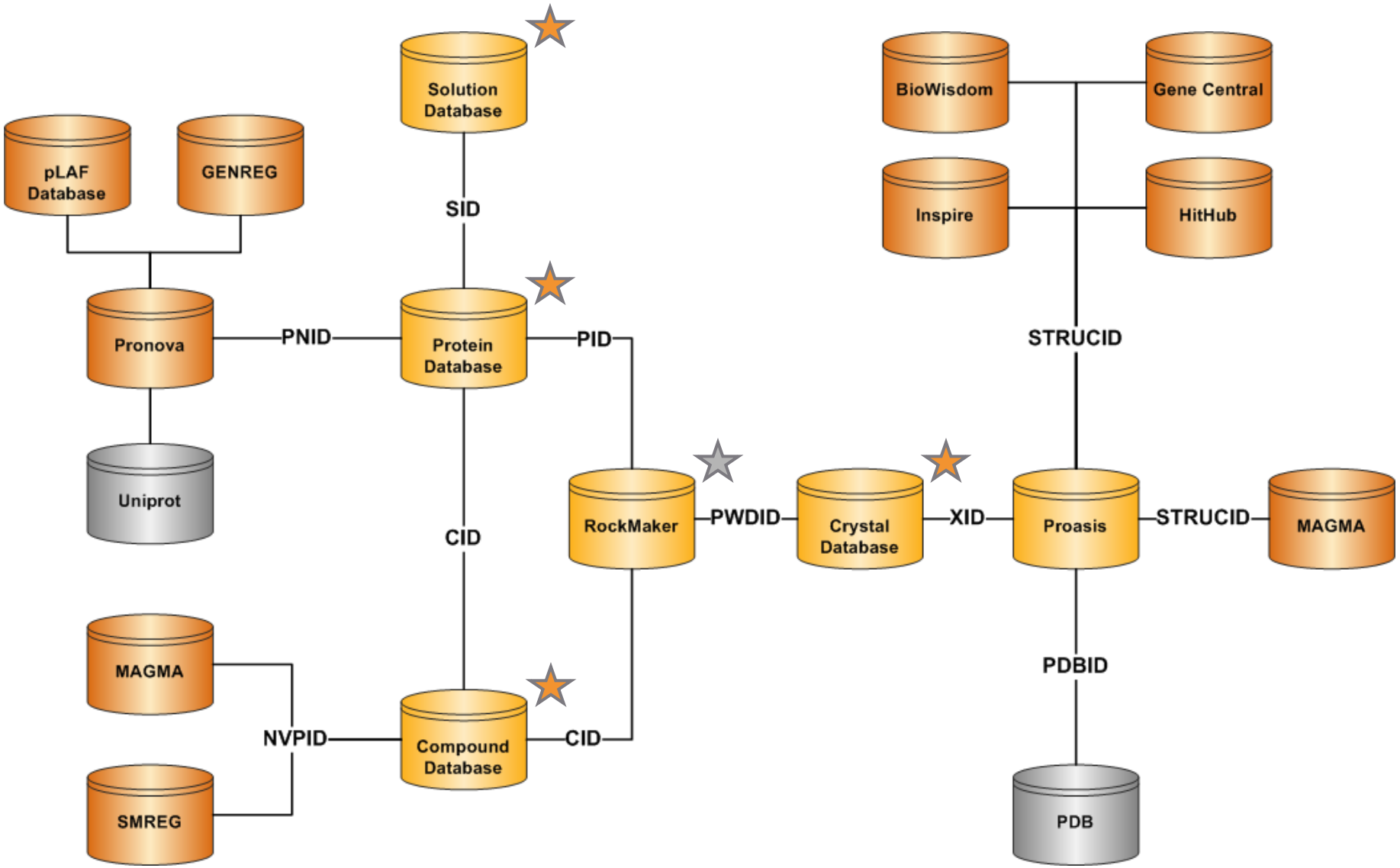
How you can make a difference

- 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

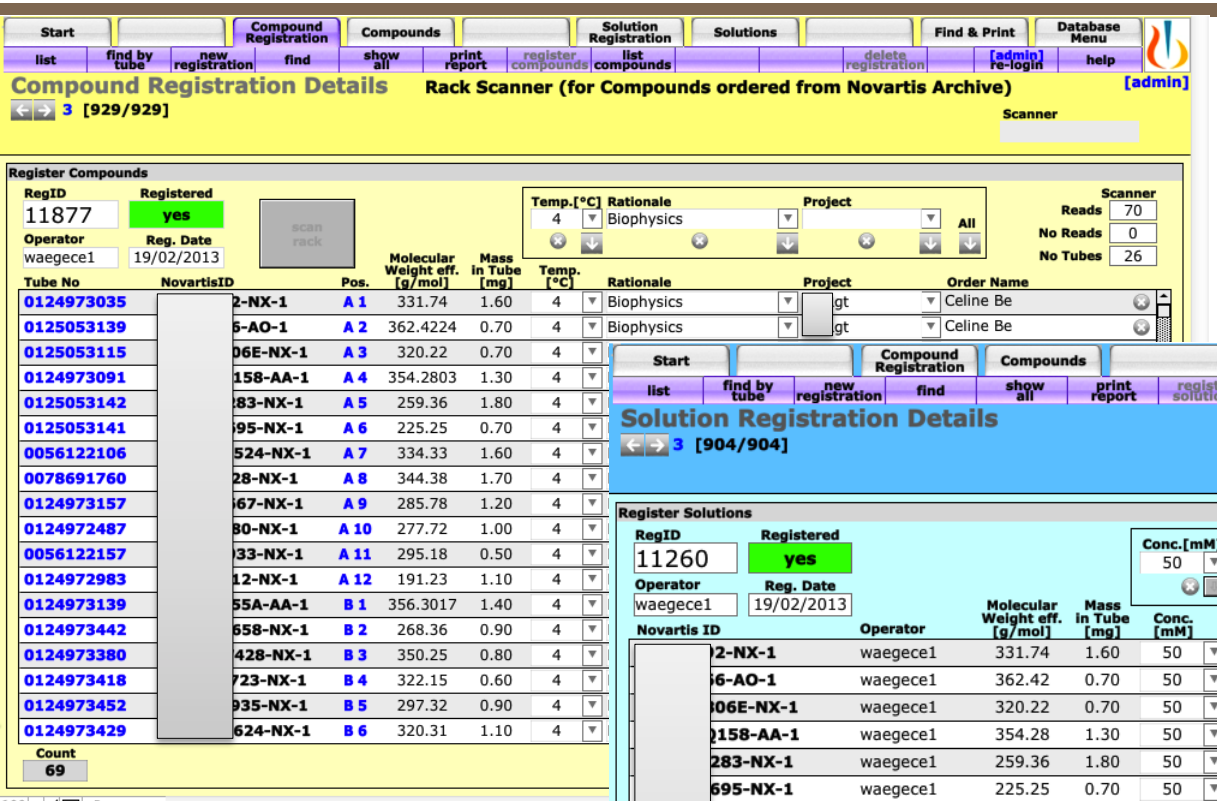
In-Process Databases

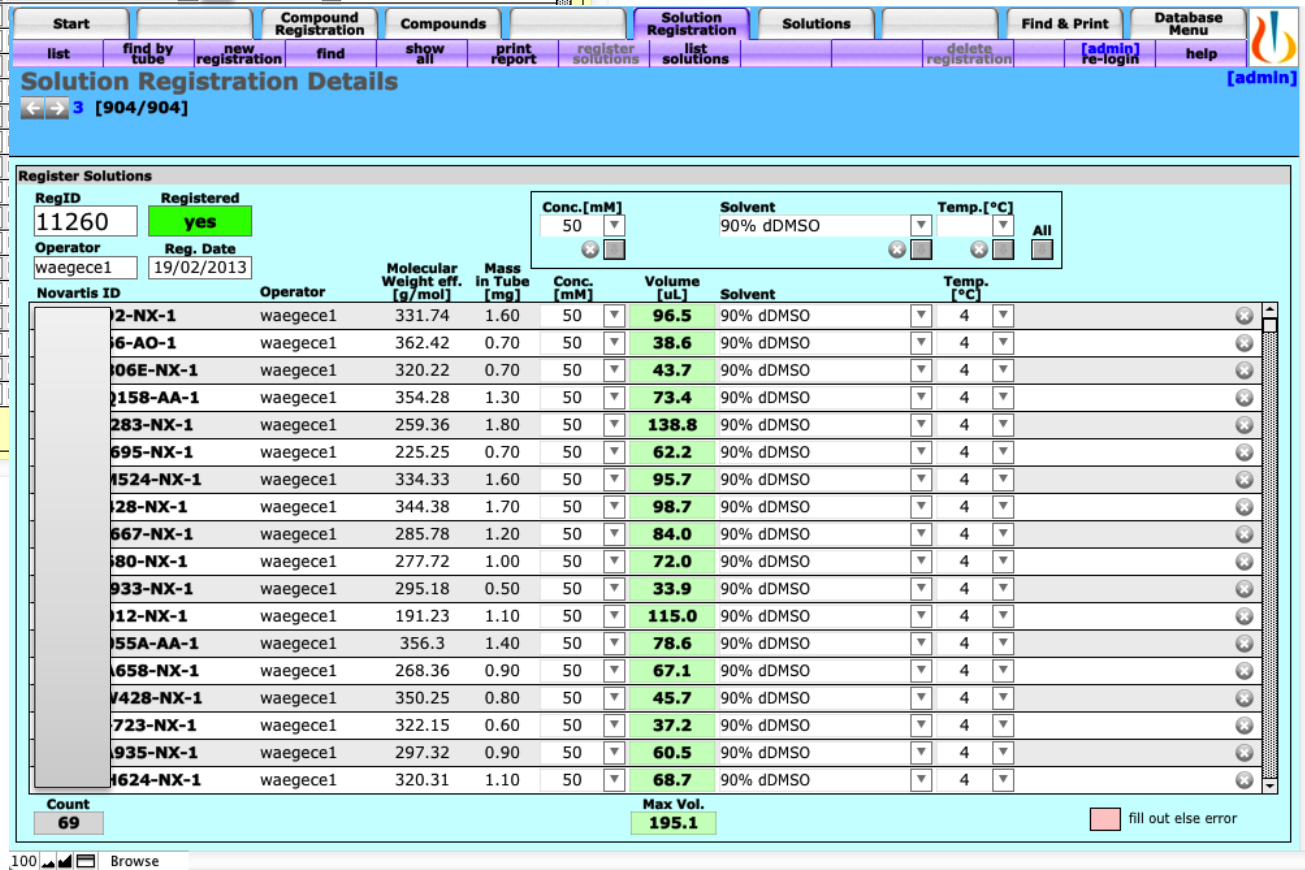
- Providing benefit to depositors
  - Scientists wants „in-process” data capturing tools which helps them to be more efficient and which are used to seamlessly upload data into end-point databases

Data Management Concept

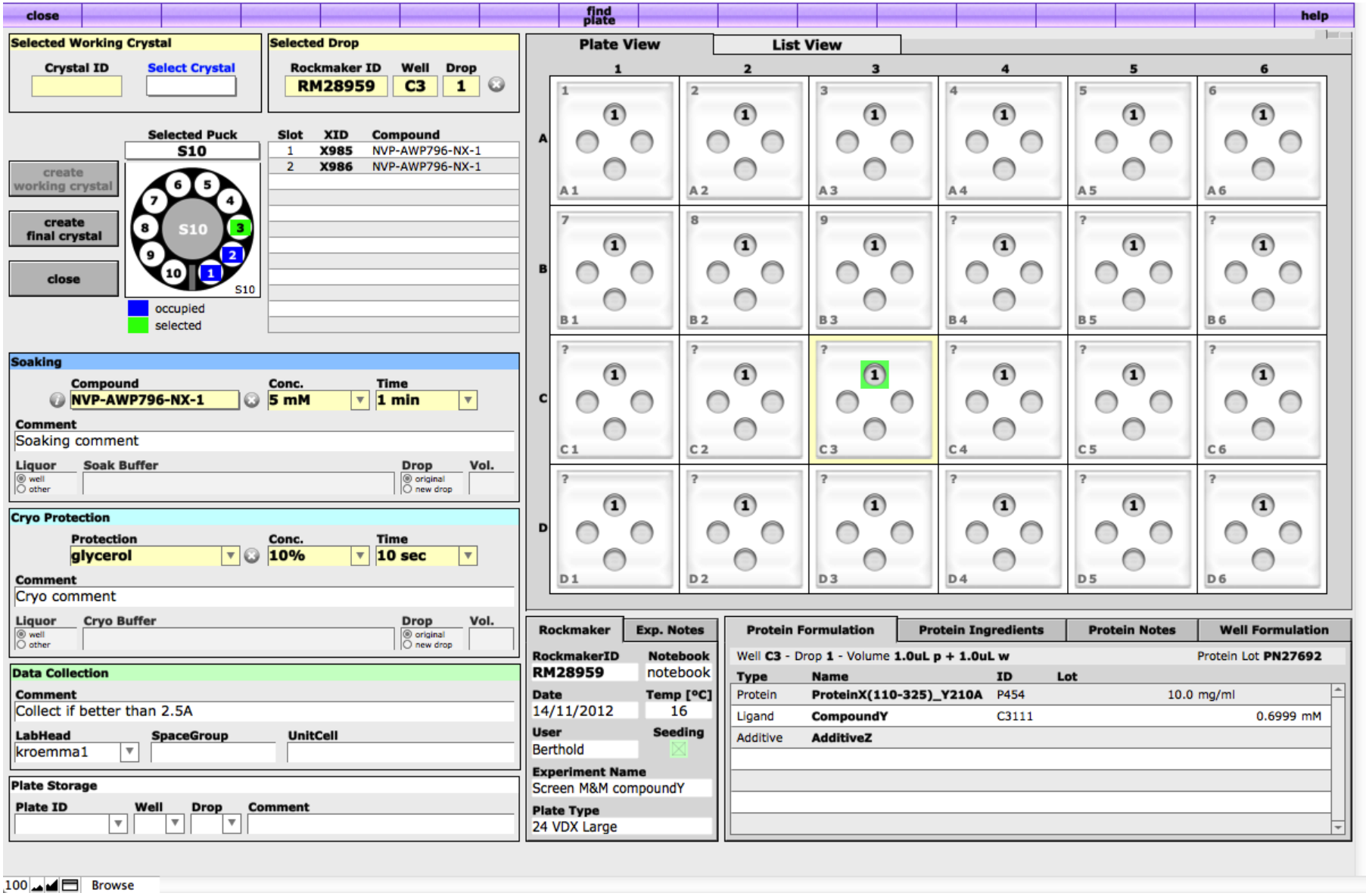


Compound Database

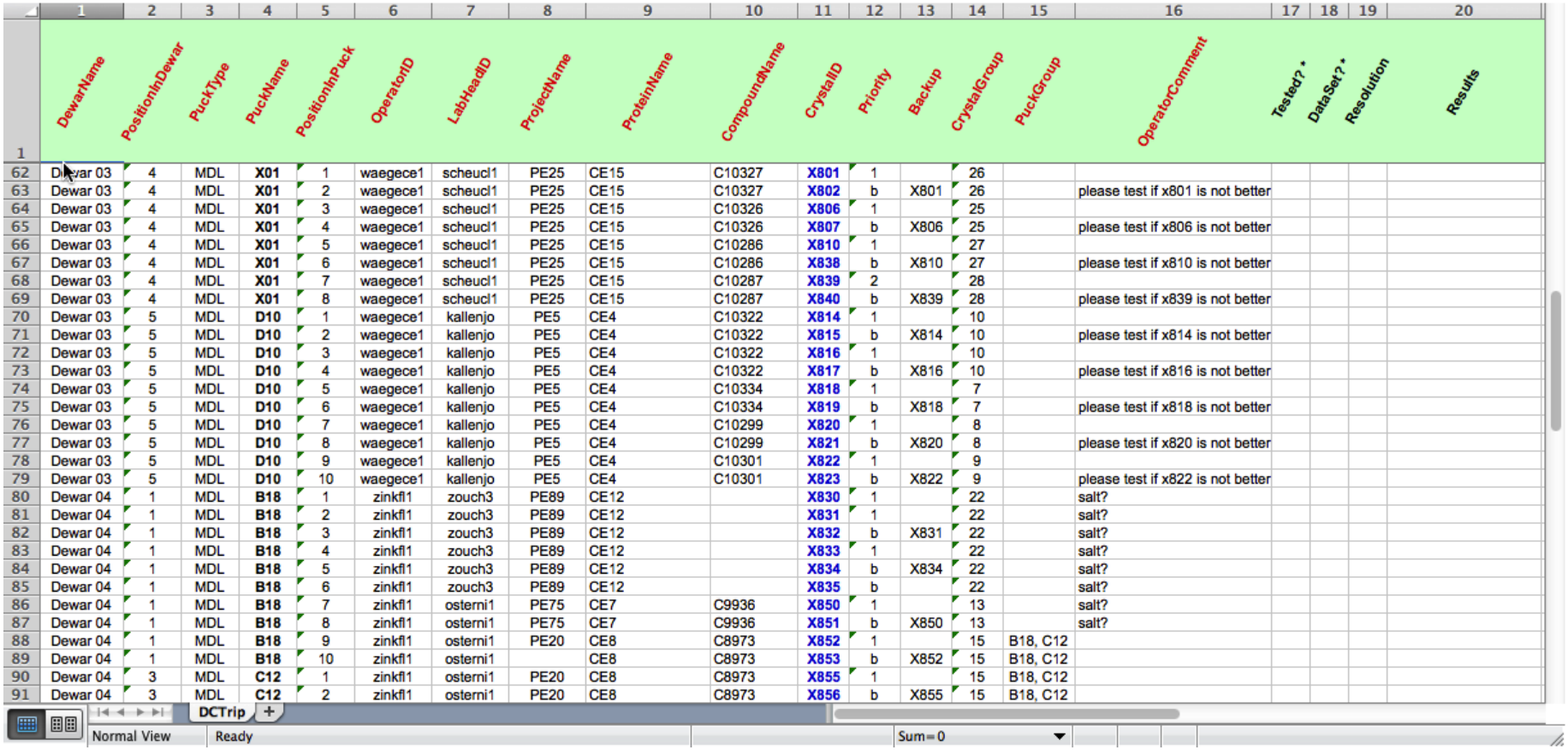




Crystal Database



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
1	DewarName	PositionInDewar	PuckType	PuckName	PositionInPuck	OperatorID	LabHeadID	ProjectName	ProteinName	CompoundName	CrystalID	Priority	Backup	CrystalGroup	PuckGroup	OperatorComment	Tested? *	DataSet? *	Resolution	Results
62	Dewar 03	4	MDL	X01	1	waegece1	scheucl1	PE25	CE15	C10327	X801	1		26						
63	Dewar 03	4	MDL	X01	2	waegece1	scheucl1	PE25	CE15	C10327	X802	b	X801	26		please test if x801 is not better				
64	Dewar 03	4	MDL	X01	3	waegece1	scheucl1	PE25	CE15	C10326	X806	1		25						
65	Dewar 03	4	MDL	X01	4	waegece1	scheucl1	PE25	CE15	C10326	X807	b	X806	25		please test if x806 is not better				
66	Dewar 03	4	MDL	X01	5	waegece1	scheucl1	PE25	CE15	C10286	X810	1		27						
67	Dewar 03	4	MDL	X01	6	waegece1	scheucl1	PE25	CE15	C10286	X838	b	X810	27		please test if x810 is not better				
68	Dewar 03	4	MDL	X01	7	waegece1	scheucl1	PE25	CE15	C10287	X839	2		28						
69	Dewar 03	4	MDL	X01	8	waegece1	scheucl1	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		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					

DCTrip + | Normal View | Ready | Sum=0

NOVARTIS

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

NOVARTIS