IJC as a Platform

IJC as a general purpose chemistry data management solution

Erin Bolstad

ChemAxon UGM
September 25 - 26, 2012

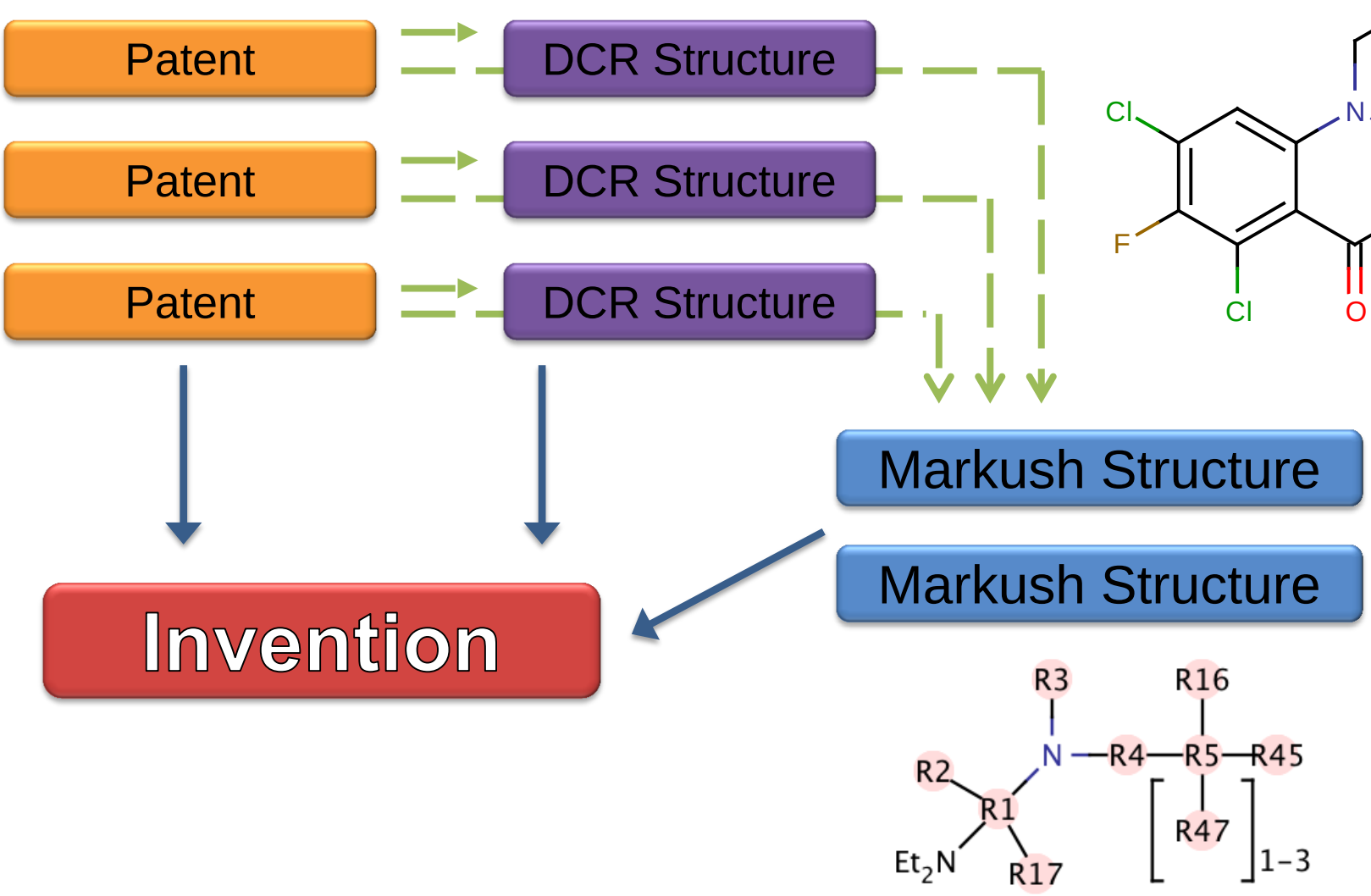
Basic IJC Platform

- Data access and manipulation
- Import/export
- Form builder
- Chemical spreadsheet
- Query execution
- Chemical terms
- …and more being added with each release

Two Ways of Extending

- Java Extension Modules
  - Much of IJC's functionality is done this way
  - Very powerful, but complex
  - You must re-deploy IJC if your extensions change
- Groovy Scripting
  - New since IJC 5.4 (Jan 2011)
  - Simpler and faster, but less powerful
  - Runtime deployment and execution

Examples: TR Patent DB

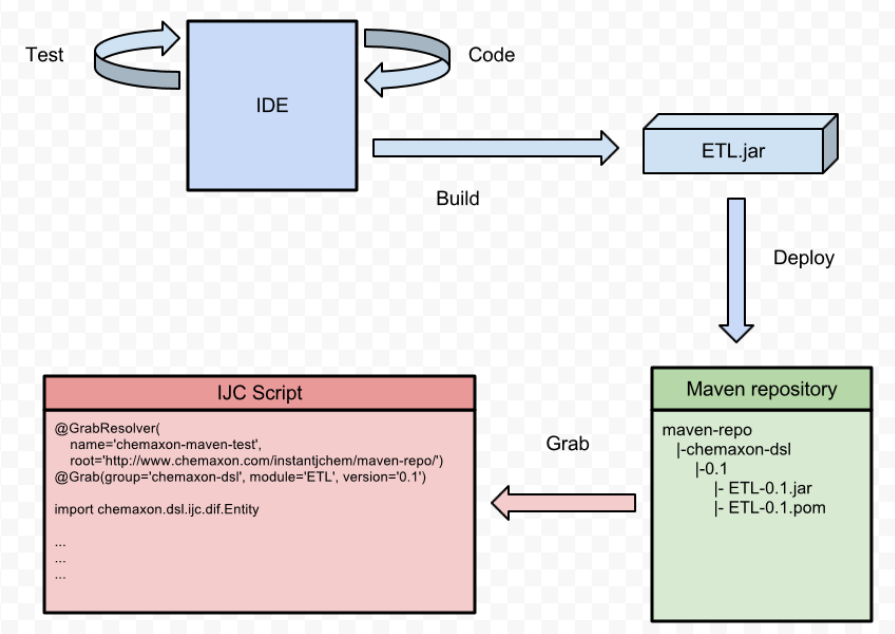


Event Model Hooks: 5.11

- Double click
- In Development
  - Selection event
  - Insert/query fields as drop-down lists
- Future Development
  - Right click -> Menus
  - Customized add/delete row interface
- Suggestions wanted!

Extensibility: Libraries

Using IJC to access external libraries

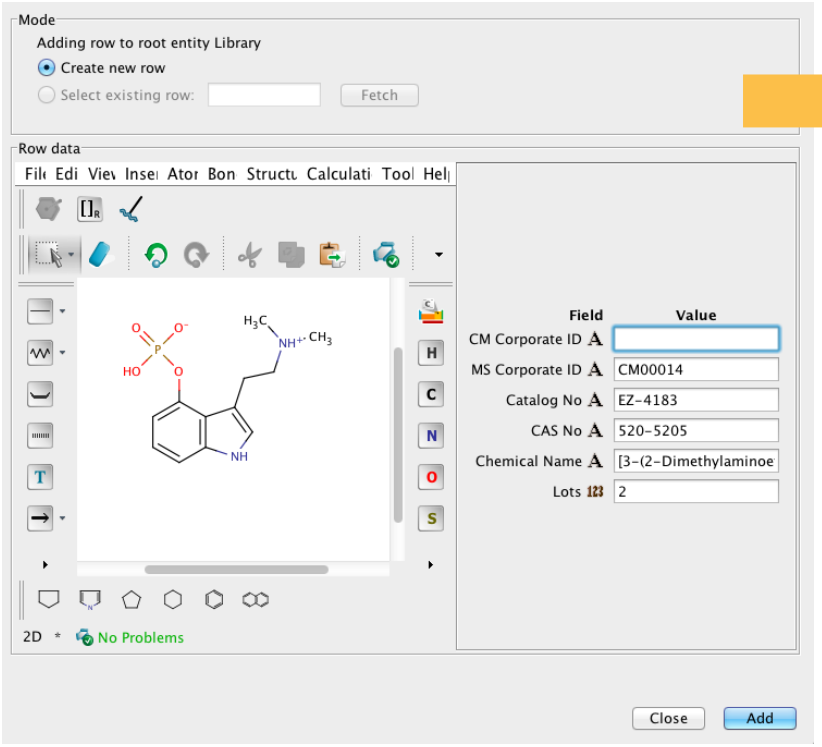


- Benefit of coding in your own IDE
  - Code completion
  - Unit testing
  - Version control
- Build library (jar file)
- Deploy to Maven/Ivy
- ‘Grab’ using Grape
- Use

Mount Sinai The Mount Sinai Hospital

Case Study: MSSM

- Needed a better way to register compounds, track data, handle spectra.



Database

(single table)

Manually assigned compound ID:

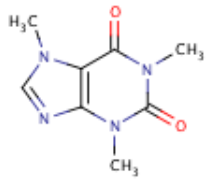
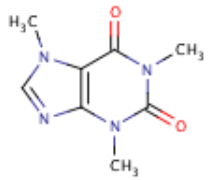
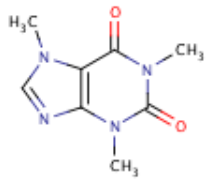
CM000137 *and/or*

MS002864

- Duplicate structures
- No easy way to analyze data
- Difficult to track lots
- Spectra in unassociated folder somewhere

Mount Sinai The Mount Sinai Hospital

Case Study: MSSM

	CdI	Structure	Corporate ID	CAS No	Chemical Name	Mol Weight	Catalog No	Formula	Location	Lot
1	1		CM0000386	58-08-2	1,3,7-trimethyl-1H-purine-2,6(3H,7H)-dione 3,7-dihydro-1,3,7-trimethyl-1H-purine-2,6-dione	194.19	C0750-5G	C8H10N4O2	Freezer 18, 32:2a	
2	3		MS008762	58-08-2	1,3,7-trimethyl-1H-purine-2,6(3H,7H)-dione 3,7-dihydro-1,3,7-trimethyl-1H-purine-2,6-dione	194.19	C0750-5G	C8H10N4O2	Freezer 23, 6:15	
3	4		MS008762	58-08-2	1,3,7-trimethyl-1H-purine-2,6(3H,7H)-dione 3,7-dihydro-1,3,7-trimethyl-1H-purine-2,6-dione	194.19	C0750-5G	C8H10N4O2	Freezer 23, 6:16	

Mount Sinai The Mount Sinai Hospital

Case Study: MSSM

Requirements for registration:

- Automatically generate corporate IDs for new structures
- Register data with links to structure, automatically tracking lot numbers
- Track lots/registration information
- Organize and link to spectra

Mount Sinai The Mount Sinai Hospital

Case Study: MSSM

Two Stage Solution:

1. Separate data from structures into non-duplicative structure library table, and data into connected data table.
2. Create script for registering new compounds

Mount Sinai The Mount Sinai Hospital

Case Study: MSSM

Database cleaned via a script

Library

- Unique structures
- Lots counted
- Corporate IDs sorted to individual fields

Data

- One row per registration
- Corporate IDs
- Lot numbers transferred
- Connected via structure relationship

Mount Sinai The Mount Sinai Hospital

Case Study: MSSM

New Registration Process:

Set of structures
(each given an Import Name)

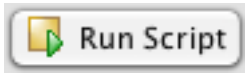
Spectra (PDFs)
(named with compound Import Name)

CSV of compound data
(connected via Import Name)

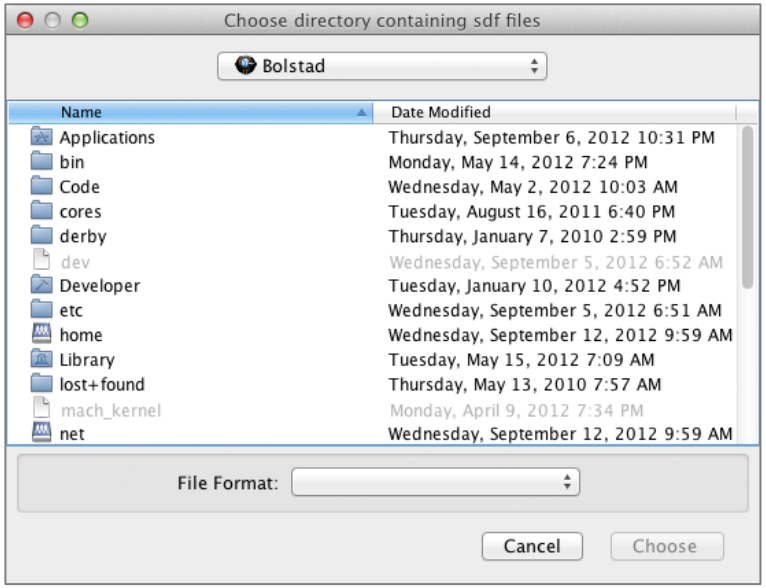
Data

Mount Sinai The Mount Sinai Hospital

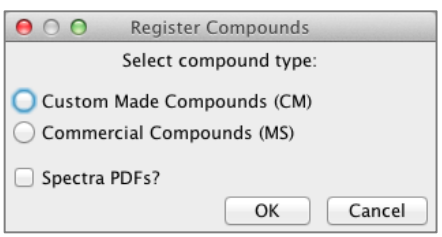
Case Study: MSSM



(Create tar of database for rollback)



Select directory of SDFs



Select ID type
Select if PDFs will be registered

Eg: (CM)

Begin iterating through directory:

For each SDF molecule

ChemAxon

Case Study: MSSM

The Mount Sinai Hospital

No → Duplicate Search

Count

	Precursor	Count
1	CM	745
3	MS	1,267

Create new Corporate ID
(CM0000745)
Increase count

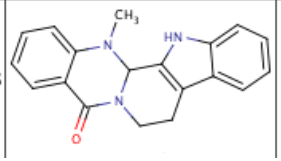
Update Data with new Corporate ID (Import Name)

Data

	ID	Lot Number	MS Corporate ID	CM Corporate ID	Import Name	Catalog No	Source	CAS
1	1			CM0000745	4GK_734B	C0750-5G	Smith4B	58-
2	2			CM0000745	4GK_734B	C0750-5G	Smith4B	58-
3	3				3N347	G78-82-126	Smith4B	518

Insert new compound with new Corporate ID

Library

	CdId	Structure	CM Corporate ID	MS Corporate ID	Lots	CAS No	Chemical Name	Catalog N
1	3		CM0000745		1	58-08-2	Evodiamine	C0750-5

Spectra()

ChemAxon

Mount Sinai The Mount Sinai Hospital

Case Study: MSSM

Duplicate Search → Yes → (CM)

Fetch all Corporate IDs

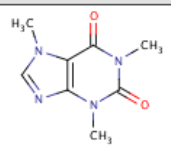
- CMxxxx → Use existing Corporate ID
- !CMxxxx → Create new Corporate ID

Update Data with Corporate ID(s) (Import Name)

	ID	Lot Number	MS Corporate ID	CM Corporate ID	Import Name	Catalog No	Source	CAS No
1	1		MS0001267	CM0000746	EZ198ba	AP10005443	Wil6P	2438-32-
2	2		MS0001267	CM0000746	EZ198ba	AP10005443	Wil6P	2438-32-
3	3				3N347	G78-82-126	Smith4B	518-17-2

Data

Update lot number, (new corporate ID) in Library

	CdId	Structure	CM Corporate ID	MS Corporate ID	Lots	CAS No	Chemical Name	Catalog I
1	1		CM0000746	MS0001267	3	58-08-2	1,3,7-trimethyl-1H-purine-2,6(3H,7H)-dione 3,7-dihydro-1,3,7-trimethyl-1H-purine-2,6-dione	AP1000

Library

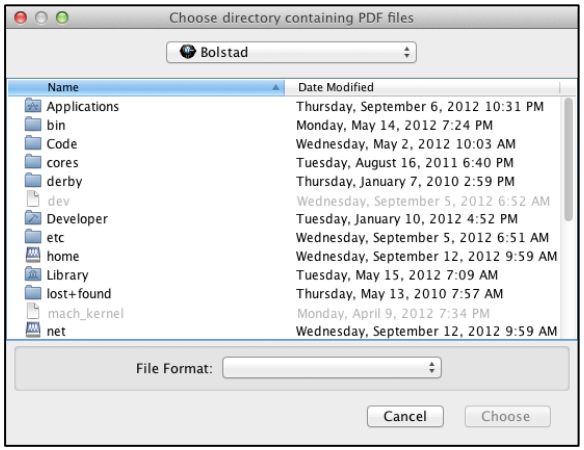
Spectra()

ChemAxon

Mount Sinai The Mount Sinai Hospital

Case Study: MSSM

Spectra



Begin iterating through directory:

For each PDF file (Import Name)

- Find Corporate ID based on Import Name
- Rename file
- Move to pre-defined repository
- Generate file URL
- Update Data with URL

r	MS Corporate ID	CM Corporate ID	Import Name	Catalog No	Source	CAS No	Chemical Name	PDF	Pr
1	MS0001267	CM0000746	EZ198ba	AP10005443	Wil6P	2438-32-6	1,3,7-trimethyl-1H-pu...	file:///Users/erinbolstad/CM0000746.pdf	Ju
2	MS0001267	CM0000746	EZ198ba	AP10005443	Wil6P	2438-32-6	1,3,7-trimethyl-1H-pu...	file:///Users/erinbolstad/CM0000746.pdf	Au
			3N347	C78-82-126	Smith4R	518-17-2	Fyodiamine		Au

Cleanup(): Remove all Import Name in Data

Future Directions: IJC MiniAps

- Clustering
- Screening
- Interface with external data (CheMBL, ZINC)
- Statistics, analysis
- Structure cleaning
- Analog and library generation/design
- Simple registration and inventory
- IP protection

What Next?

- Learn more at the developer training
- Look at the IJC developer docs
- Try the examples
- Collaborate with us!