ChemInfo – A Bridge Connecting Data Islands

Dr. Bernd-Jürgen Freitag
NIBR / GPS / SB

ICIC Conference, Nimes, Oct. 2006

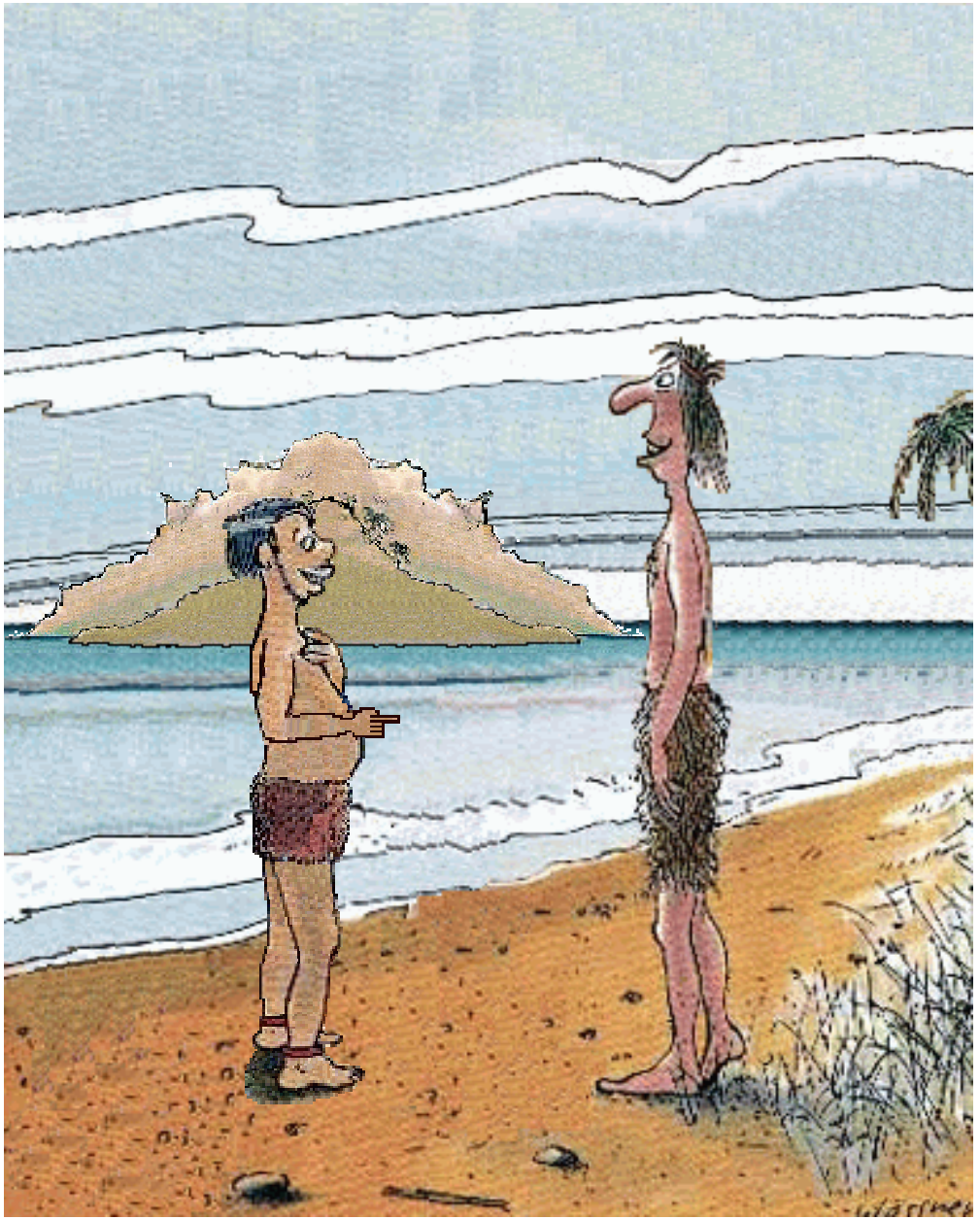
Welcome on board !

Welcome to a journey

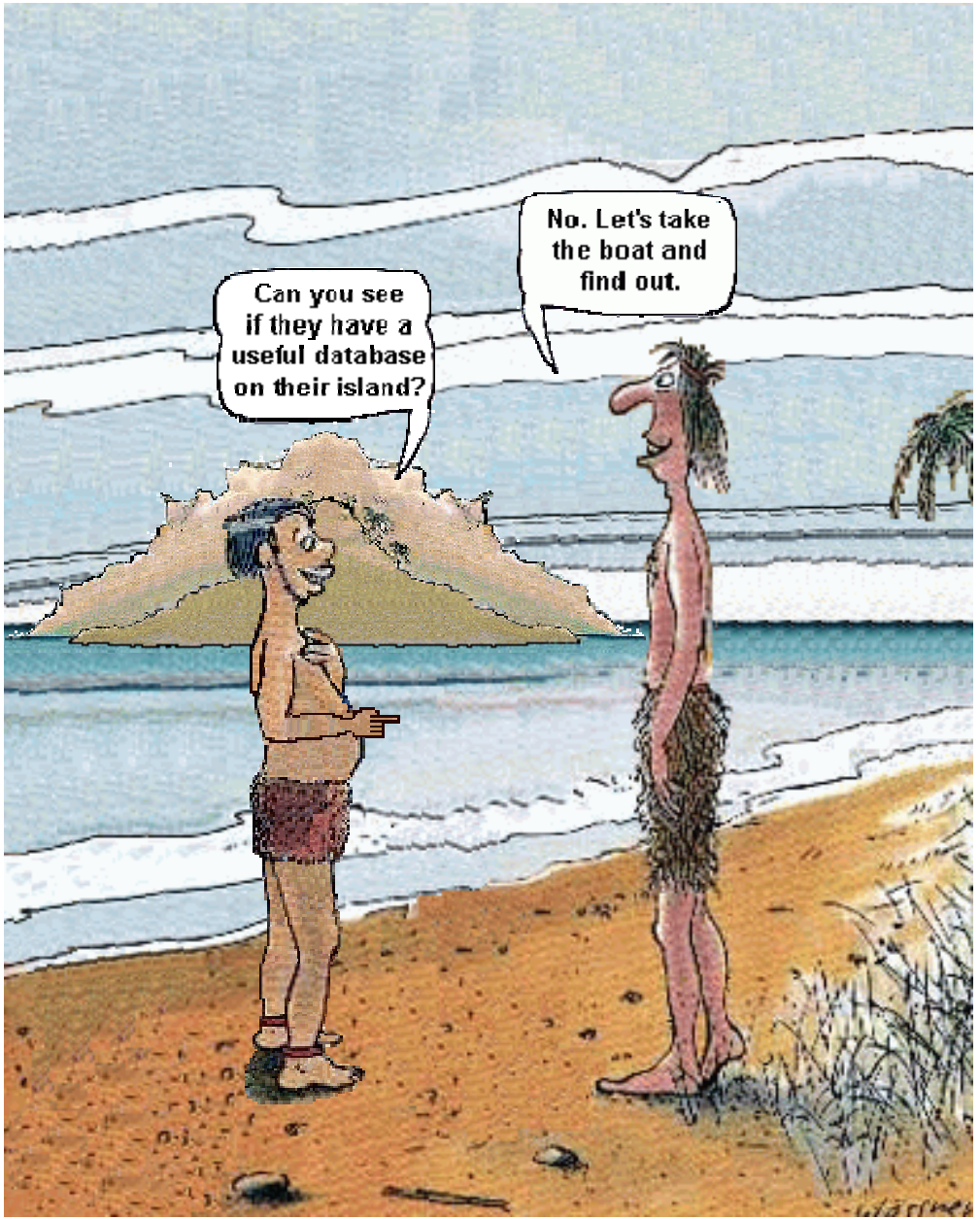
across the *mare nescentiae

to the *islands of knowledge* ...

Does this picture tell the truth ? Apparently not, because there is something missing ...



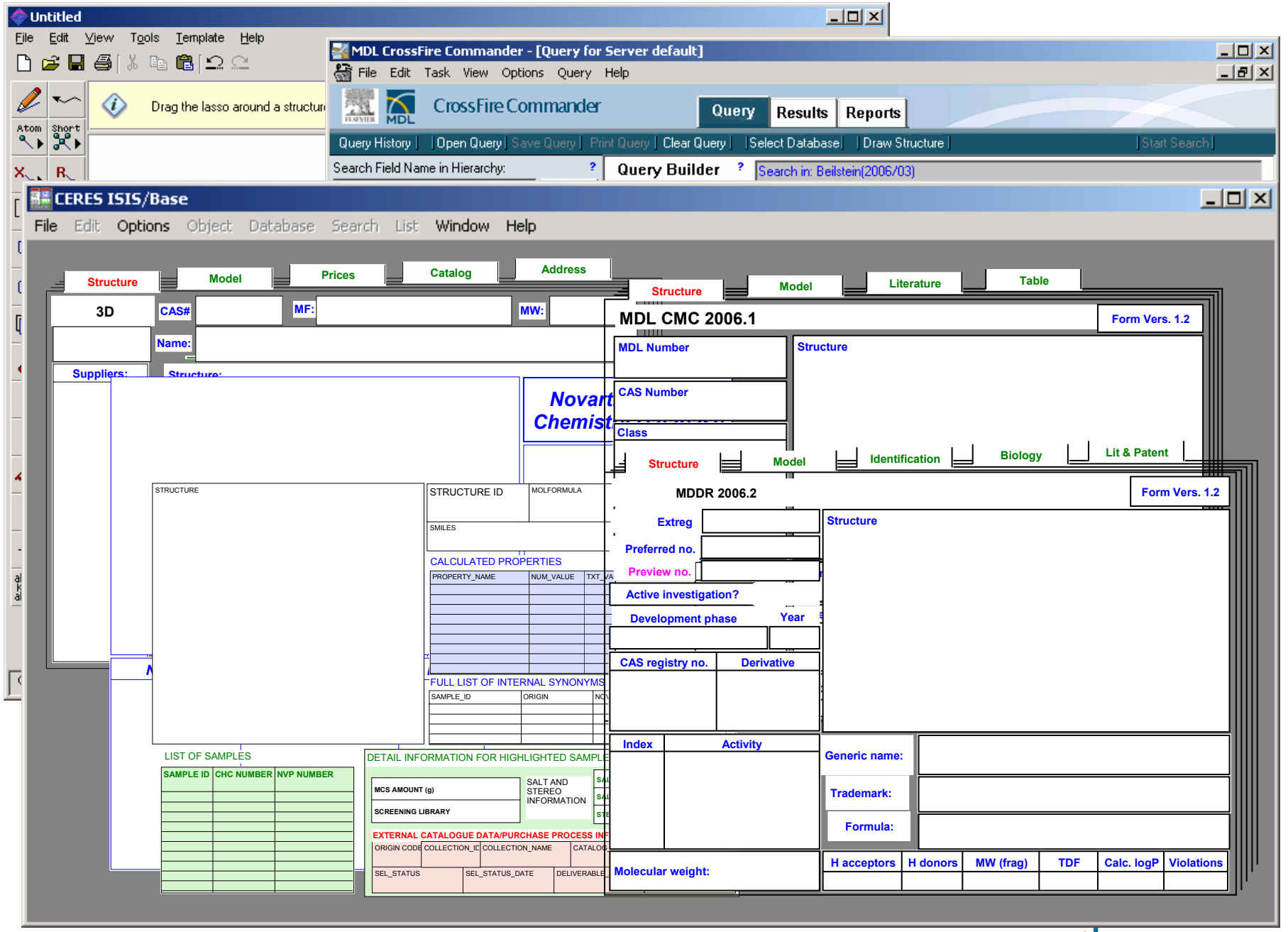
... and now we understand the situation ...


Can you see if they have a useful database on their island?
No. Let's take the boat and find out.

Even an advanced equipment didn‘t help to answer the question from distance

Application Landscape 2000

- User‘s Dream
  - Global Search over all databases through a single interface by a single query yielding a single answer set
- Reality:
  - Dedicated Applications and/or
  - Dedicated Application Windows


Untitled
File Edit View Tools Template Help
Drag the lasso around a structur
MDL CrossFire Commander - [Query for Server default]
File Edit Task View Options Query Help
CrossFire Commander
Query Results Reports
Query History | Open Query | Save Query | Print Query | Clear Query | Select Database | Draw Structure | Start Search
Search Field Name in Hierarchy:
Query Builder
Search in: Beilstein(2006/03)
CERES ISIS/Base
File Edit Options Object Database Search List Window Help
Structure Model Prices Catalog Address
3D CAS# MF: MW:
Name:
Suppliers:
Structure:
STRUCTURE
STRUCTURE ID MOLFORMULA
SMILES
CALCULATED PROPERTIES
PROPERTY_NAME NUM_VALUE
FULL LIST OF INTERNAL SYNONYMS
SAMPLE_ID ORIGIN
LIST OF SAMPLES
SAMPLE ID CHC NUMBER NVP NUMBER
DETAIL INFORMATION FOR HIGHLIGHTED SAMPLE
MCS AMOUNT (g)
SCREENING LIBRARY
SALT AND STEREO INFORMATION
EXTERNAL CATALOGUE DATA/PURCHASE PROCESS INF
ORIGIN CODE COLLECTION_ID COLLECTION_NAME
SEL_STATUS SEL_STATUS_DATE
Structure Model Literature Table
MDL CMC 2006.1
Form Vers. 1.2
MDL Number
Structure
CAS Number
Class
Structure Model Identification Biology Lit & Patent
MDDR 2006.2
Form Vers. 1.2
Extreg
Preferred no.
Preview no.
Active investigation?
Development phase
Year
CAS registry no.
Derivative
Structure
Index
Activity
Generic name:
Trademark:
Formula:
Molecular weight:
H acceptors H donors MW (frag) TDF Calc. logP Violations

Application Landscape 2000

- User‘s Dream
  - Global Search over all databases through a single interface using a single query
- Reality:
  - Dedicated Applications and/or
  - Dedicated Application Windows
- **New Approach: Data Warehouse**
  - **Allows very sophisticated queries**
  - **But requires extensive training**

Goals for ChemInfo

- **Access to all data sources with a single query**
- **Display all results as a single hit list**
  grouped by structure or each compound number separately
- **Smart application, i.e.**
  - **Few search options easily to understand**
  - **Few display options, but all clickable**
- **Fast development with very limited resources**
  - **Kick-off in February 2001**
  - **Presentation of the application at Research Conference in September 2001**

Rapid Development

- **Reduction to essential aspects of development**

  User requirements and implementation
  are the principal components of software development

  Thus they are orthogonal, says statistics

  Which means, they have nothing in common

  - **Concentration on implementation**

  Project planning + documentation demands a lot of time

  - **Concentration on implementation**

Rapid Development, continued

- **By elimination of minor aspects of development, we had enough time to implement also minor aspects of functionality**
- **Most important: User Interface (~ 50% of time)**
  weekly presentation and discussion of the actual development in our group
- **Avalon under development at that time, Network quite instable → modifications (~ 20%)**
  (Avalon = Novartis Pharma Data Warehouse)
  e.g., automatic switch to MDL DBs if Avalon failed, projection of number of hits + search time
- **Development of core functionality (~ 30 %)**

Development Phases

- **Start in February 2001**
  Avalon (Novartis Warehouse, Pharma + Agro + ACD + WDI), MDDR + CMC as MDL databases
- **Beta version presented to power users in June 01**
  - **New request:**
    Provide manual selection as ISIS does
  - **Implemented in four weeks**
    e.g., invention of a *checkbox with 3 states*

- **Beilstein integrated in four weeks**
  e.g. MDL Molfile – Rosdal conversion for query options
- **First Release at Research Conference, September 01**

ChemInfo Client in Few Words

- Tabs for StrQry, DataQry, Result, Browse, Lists
- Two query forms
  - Structure Query
    - StructureBox for exchange with ISIS/Draw
    - list of databases
    - search type selection (FSS, SSS, Similarity)
  - Data Query
    search terms not related to datafields
    but rather to categories of datafields
    - big box for any kind of compound code/name
      incl. CAS Reg.No., many items at once
    - text boxes for Novartis Pharma specific fields (single value)
    - button for special function Get Substance Sheets
    - button to search all databases, all equiv. numbers included

ChemInfo Client in Few Words, continued

- **Result Overview**
  - Query with options (Structure query: image & molfile link)
  - Summary for all selected databases
    for (de)selection of databases for display
  - Selection for browse layout, button to apply it
    - two layouts for display grouped by structure
    - one layout to display structure with essential data using data source specific layouts
  - Button to get all database keys / compound numbers of actual selection for transfer to Excel
  - Selector to change display of manual selection (selected, deselected, all)

ChemInfo Client in Few Words, continued

- **Browse layouts**
  - Concept of pages (should fit on a printed page)
  - Buttons to move through pages
  - Buttons for manual (de)selection
    (per page, all of current set)
  - 1st layout shows only first compound number + structure
    (grouped by structure)
  - 2nd layout: max. 7 compound numbers + structure
    (grouped by structure)
  - 3rd layout: structure + essential data
    (6 items per page, ungrouped)
  - Checkbox for display & change of selection state
  - Display of substance sheets for all related compound nos.
    (all information from a data source) in case of layout 1 & 2
  - A click on a structure image sends the molfile to ISIS/Draw

ChemInfo Client in Few Words, continued

- **Lists**
  - Overview on all hit lists of last month
  - Redisplay of the query used for the hit list
    - Transfer of a molfile to ISIS/Draw
  - Get a hit list as actual hit list
  - Save a hit list & query permanently
  - Reuse a permanently saved hit list or query

Usage

- **Fully productive release December 2001**
  incl. Beilstein Hitlist Viewer to display Beilstein details
- **Strong increase in usage during 2002**
- **Now ~ 800 users, many are infrequent users (450-500 users per quarter)**
- **> 35,000 queries during last 12 months**
- **Effort ?**
  - **< 10 demos since 2001 (this one included)**
  - **not a single training session at all**

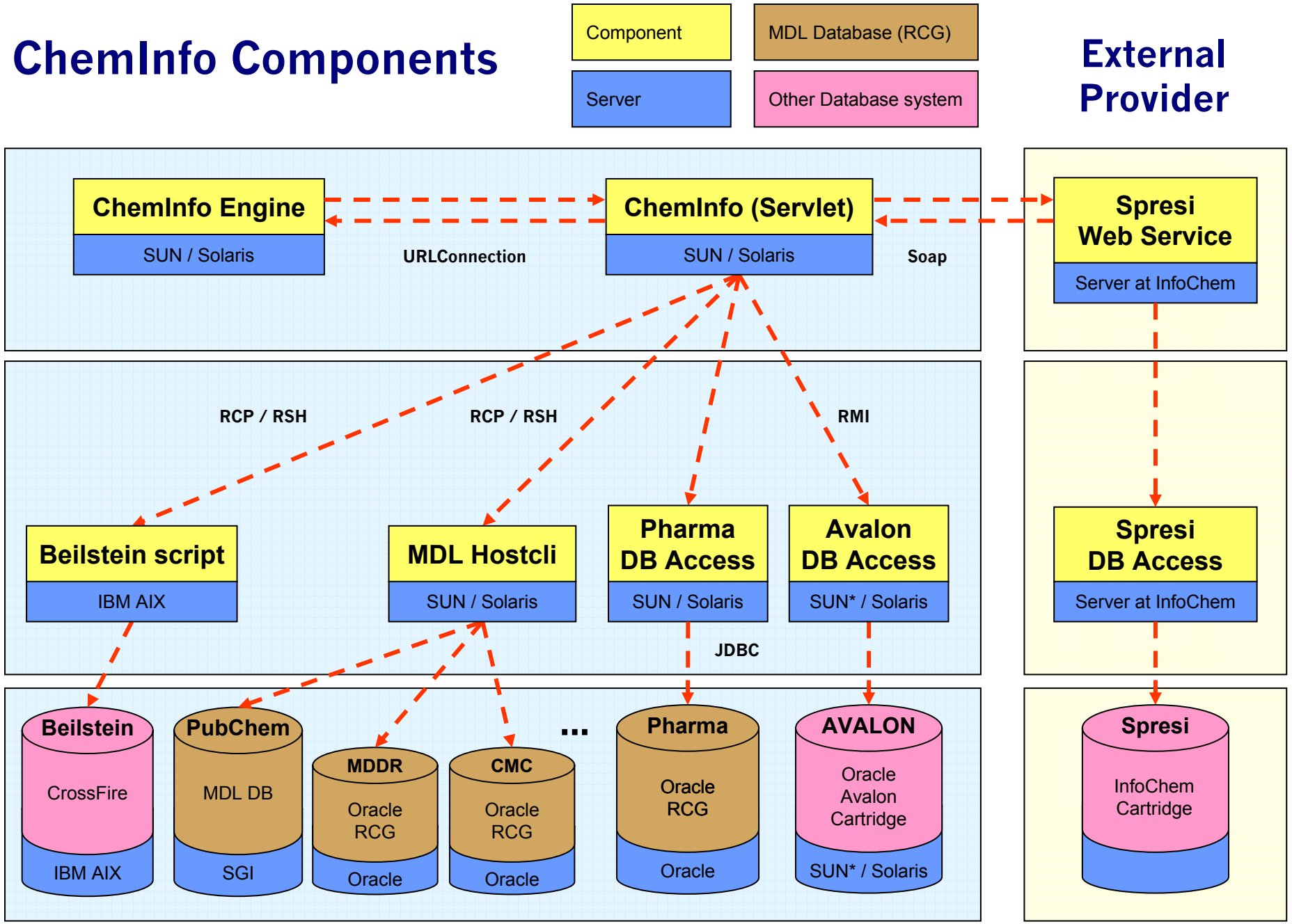
Architecture

- Multi-tier application
  - CGI / Perl for business logic + rendering
  - Dedicated database connectors
    - shell scripts + command line interface
    - Java / RMI component for Avalon
    - sqlplus (later Java / JDBC) to access Oracle
- Web browser as user interface
- ActiveX control for transfer ISIS/Draw ↔ web
- Separation of display (image) / handling (mol)
  - no plugin for display but image with link to molfile

Further Development since 2002

- **Integration of further data sources**
  - **10 in productive version (17 million structures),**
  - **15 in development version (32 million structures) incl. PubChem (inhouse), Spresi (external web service)**
- **Java servlets** for managing + performing searches and data collection, access to mol files, image generation
- **Parallelization of searches**
- **Display of progress + already retrieved hits, rendering of a temporary result** (not yet released)
- **Integration of an external web service (Spresi database, InfoChem) in a demo version**
- **Not yet done: direct access via cartridges**
  except Avalon cartridge for Avalon Data Warehouse

ChemInfo Components



NOVARTIS

Demo

- classical ChemInfo client

- prototype including an external web service

Conclusion

ChemInfo has proven as an easy-to-use integration tool for the various data bases we have at Novartis.

The integration of Spresi web service was my *Hello World* to web services offered by external information providers.

Now it‘s time for the world to answer.

I would like to encourage information providers to offer web services which allow a deep integration of information at customer side.

ChemInfo is prepared for this.

A special thank-you to Peter Löw (InfoChem) for providing the Spresi web service

Acknowledgements

Novartis

- Matthias Hübscher
  Bernhard Rohde
  - Java RMI to Avalon
- Dieter Stahel
  Willi Sieber
  Urs Hegi
  Matthias Hübscher
  Jürgen Grob
  - Review of User Interface
- Bernhard Rohde
  Andreas Saur
  Andrew Hanson
  - StructureBox ActiveX

MDL Beilstein

- Helmut Grotz
  André Bahl
  - Beilstein Commandline Interface
  - Rosdal – Molfile converter

InfoChem GmbH

- Heinz Saller
  - Java classes
    for Soap web service