

# Bridging knowledge gaps in drug discovery with semantic technologies

---

**Conference on Semantics in Healthcare and Life Sciences (C-SHALS)**

Therese Vachon, Head of Text Mining Services

Novartis Institutes for Biomedical Research, NITAS

Cambridge/Boston MA, February 26 2009

# Agenda

---

- Challenges
- Strategy
- Aggregation of Text Mining Services Activities
- Knowledge Integration Framework
- Answering Complex Queries
- Introducing a Smart Search on Research Data
- Automatic Population of specialized Wikis
- Patent and Literature Mining

# Challenges and requests

---

- Knowledge Integration / Data Federation
  - Complex drug discovery questions require federation of proteomics, genomics, and structural databases
  - The use of ontologies is mandatory to federate data sources and to elucidate, model and share knowledge about chemical, biological and disease mechanism information
  - We need to develop an environment which will foster collaboration and communication
  - Proprietary data needs to be curated, annotated and mapped to other sources
- Mining of chemical and biological knowledge contained in e.g. patents and literature

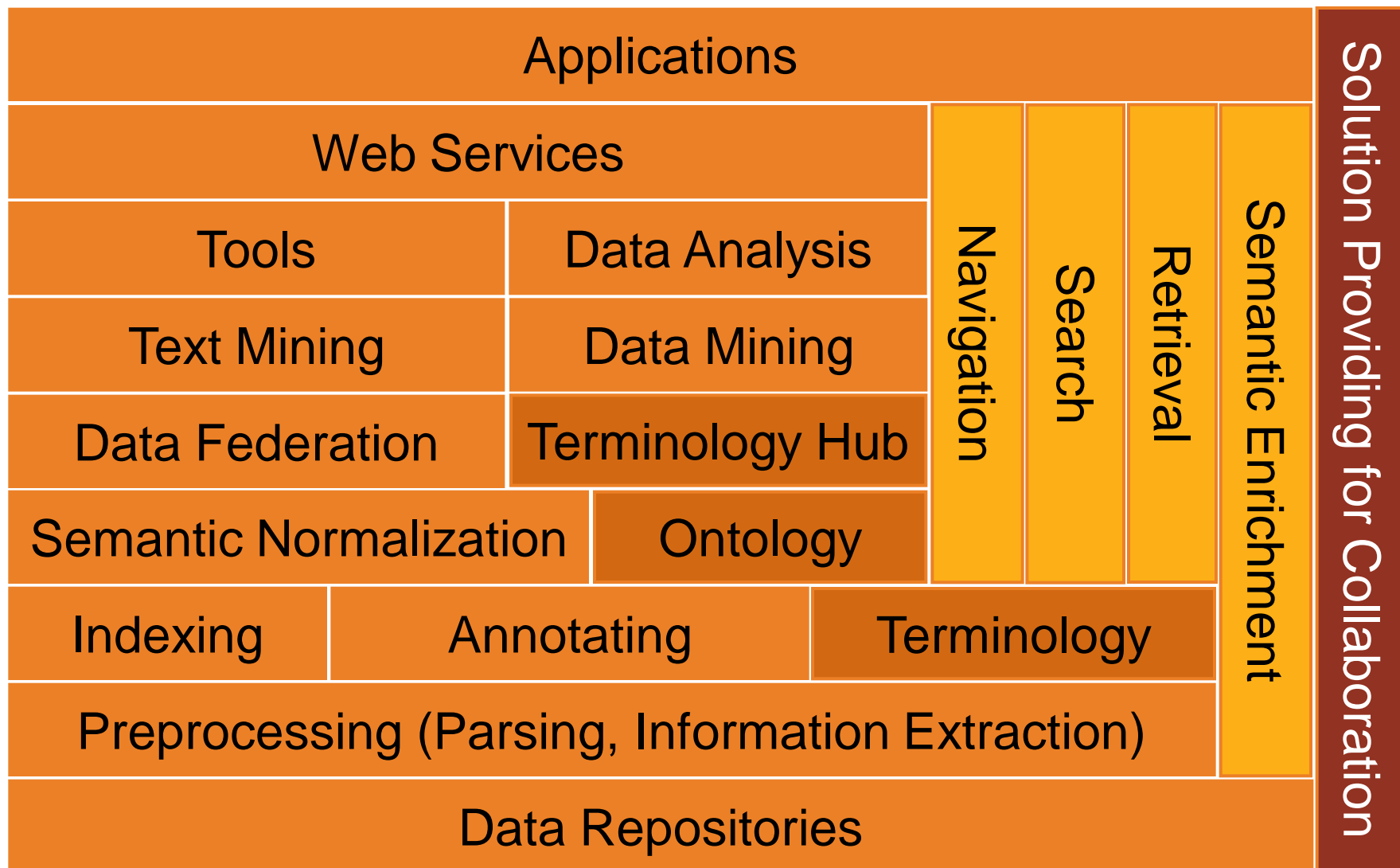
# Strategy

---

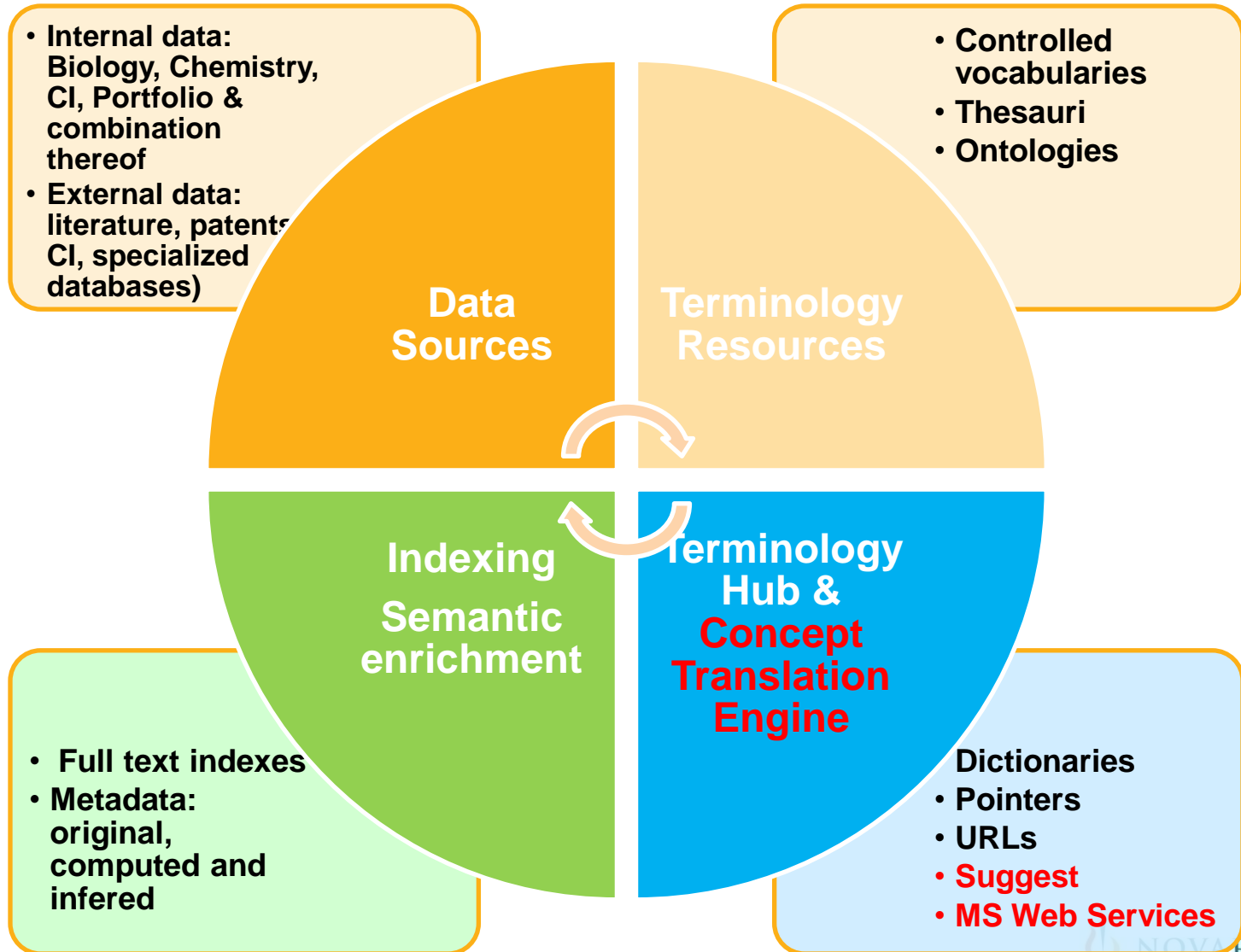
- Implement text mining solutions and a framework for data / knowledge integration that will enable researchers to
  - derive high quality information from texts and data
  - retrieve, link, synthesize, analyze, infer and interpret Life Sciences data.
- Iterative approach with experts in each area: from problem identification to the solution.

# Overview and Aggregation of Activities

## *The Text Mining Services Stack*



# Knowledge Integration Framework Concept Translation Engine



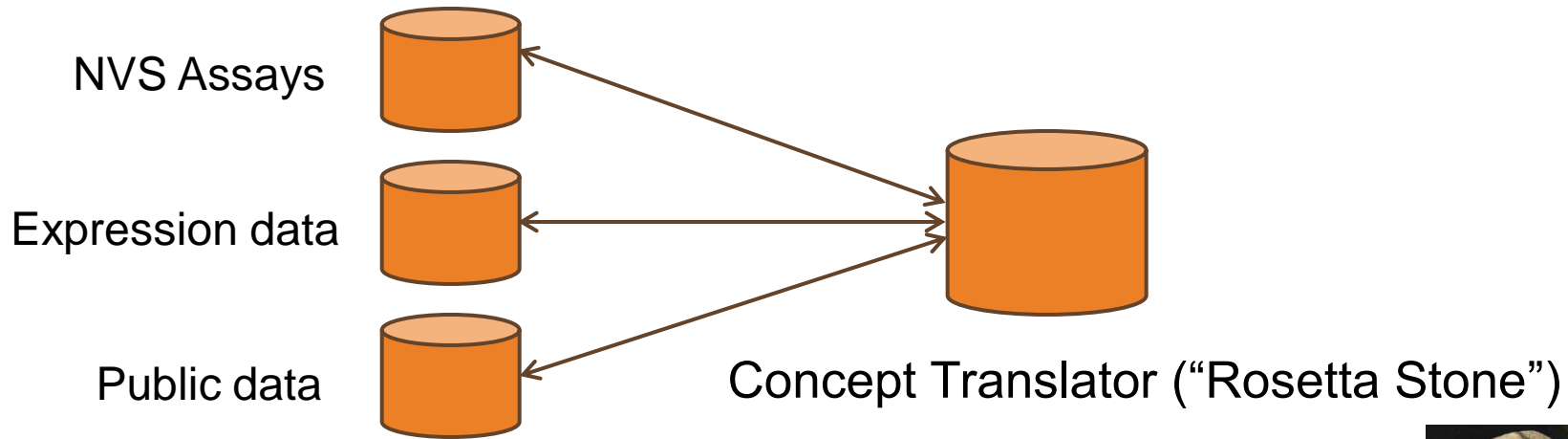
# Answering complex queries / Use Case

---

Question: Are there other kinases in pathway X for which we have potent chemical inhibitors ( $IC_{50} < 1 \text{ uM}$ ) and what cell lines could be used for activity validation ?

1. Find all proteins in pathway X
2. Subset by kinases
3. Find  $IC_{50}$  assays that measure these kinases
4. Subset by  $IC_{50}$  value  $< 1 \text{ uM}$
5. Find gene expression data for cell lines
6. Get expression values for the kinase subset
7. Return list of compounds, their kinase target and cell lines where the kinase is expressed

# Answering complex queries / Use Case



Gene <-> Pathway

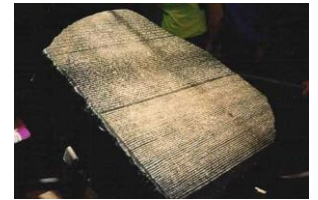
Gene <-> Family

GeneID <-> Source ID (Assay ID)

Assay type (cellular assay?)

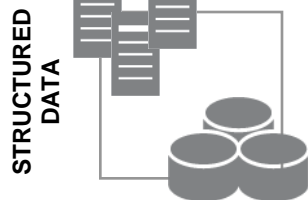
GeneID <-> Cell Lines

Filter for Gene Expression and IC50



# Applications

## NIBR Search



People



emails



### Unstructured data

NIBR SharePoint, Wikis, Blogs  
NIBR Intranet, Extranet, Internet  
NIBR G Drives

*Enterprise Search will change  
the way people interact  
– with information, people, and applications*



Search  
will be  
everywhere



Search will  
enable unique  
user experiences



Search will  
change the way  
people do  
business

**400 formats supported among those** fast

Web pages (e.g. XML, HTML)  
Files, documents (e.g. MS Office, PDF)  
Database content (e.g. Oracle)  
Applications (e.g. Exchange)

### Unstructured data

XML Feeds  
Literature, CI, Patents,  
News, ...  
Other Intranet, Internet



Research  
documentation  
Preclinical Safety, ...

### Structured data

Legacy data  
(Relational data)

### People

Novartis people  
Communities

### Emails

Exchange

# Collaboration with Business Organizations

## *Construction of specialized Wikis - Rationale*

---

- Collate relevant information to enable decision making
  - Data from different sources can be readily integrated
  - Solidify expected links, create unexpected links
  - Break down information silos
  - Capture expertise
  - Track early targets, track compounds, ...

# Collaboration with Business Organizations

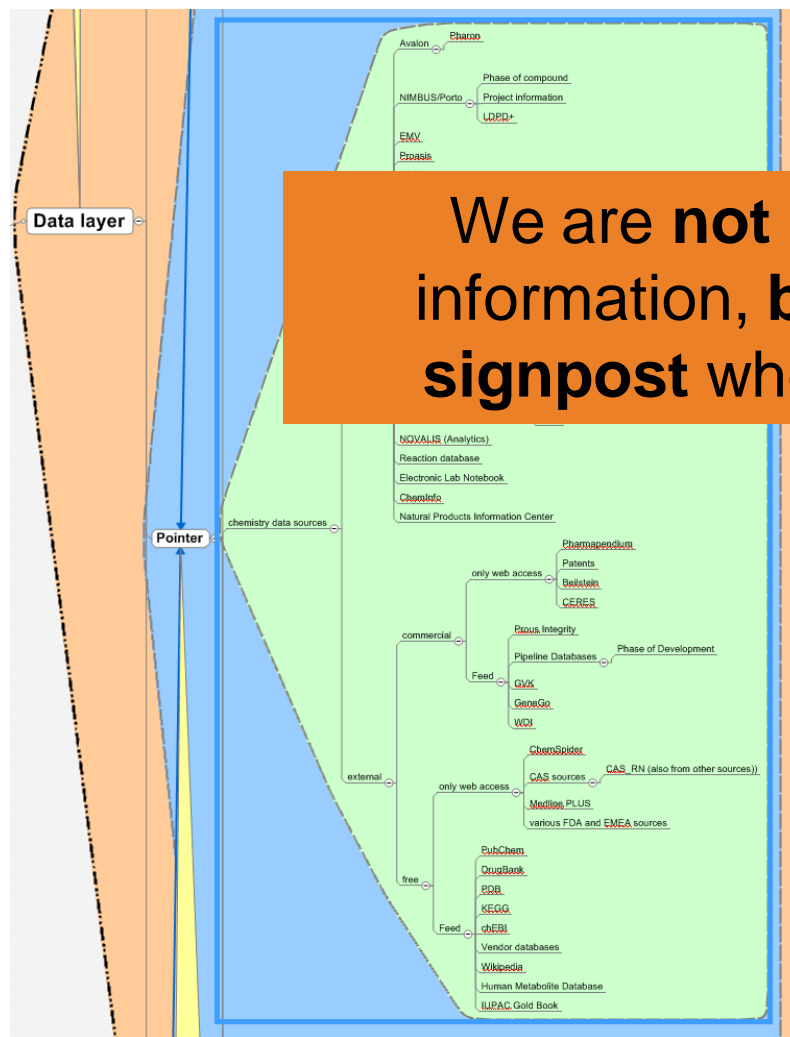
## *Construction of specialized Wikis - Rationale*

---

- Provide a collaborative authoring environment
  - Wikipedia style group contributions form natural collaborative teams and encourage participation
  - Organizing principles provide multi-faceted views of information
  - Flexible views can be created by users
- Web 2.0/3.0 architectures
  - Increase the power and flexibility of the framework to serve and integrate data

# Population of NIBR Wikis, example on Chemistry

*Widely scattered chemistry data sources*

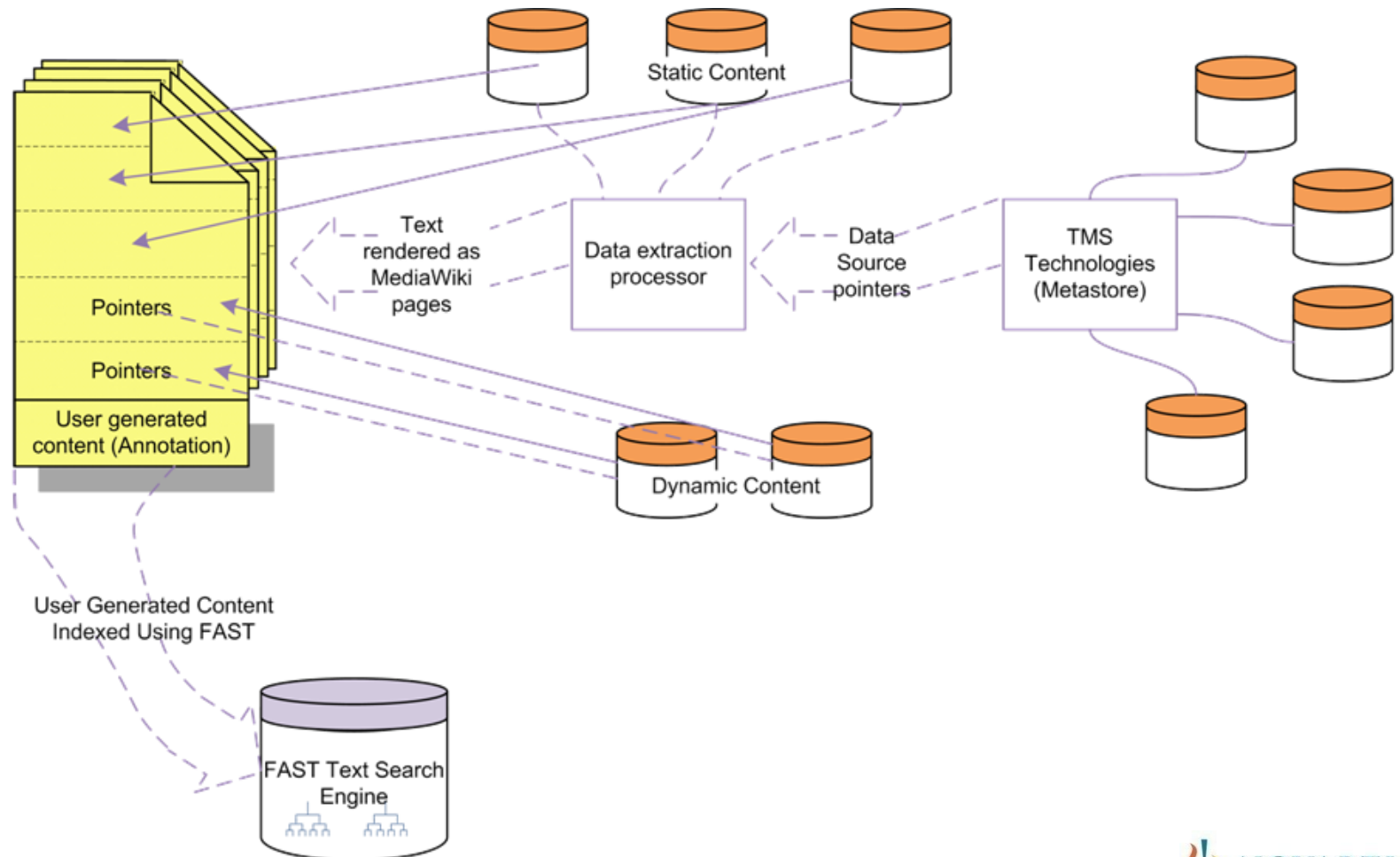


**We are not replicating information, but provide a signpost where to find it.**

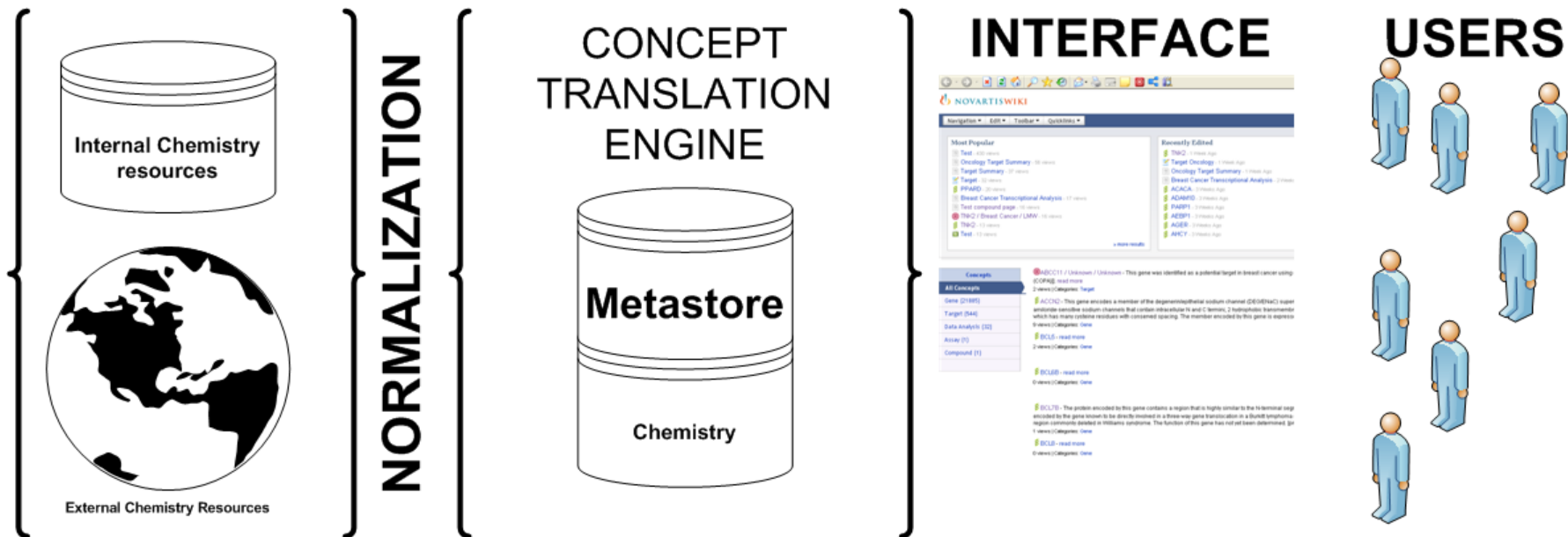


External Chemistry Resources

# Using the wiki as a framework *to view collated annotation*

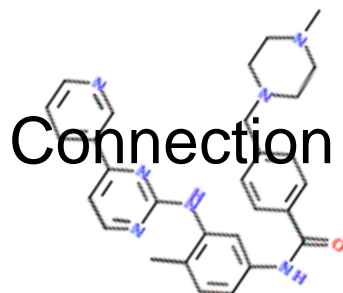


# Concept (Snapshot for Wiki Only)



# Mining for Chemical Knowledge - Rationale

- Make text corpora searchable for chemistry (e.g. Project Prospect/RSC)
- Generate chemistry databases for use in research based on Scientific Papers or Patents
- Link Chemical Information with further annotation in an automated way for e.g. Chemogenomics applications (IBM ChemVerse, Chempider)
- Patent analysis for MedChem projects

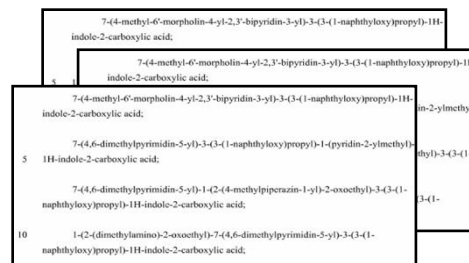


# Mining for Chemical Knowledge – Use Case

Medicinal Chemist wants to synthesize competitor compound as tool compound for own project



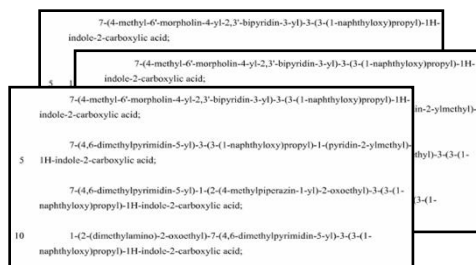
Patents



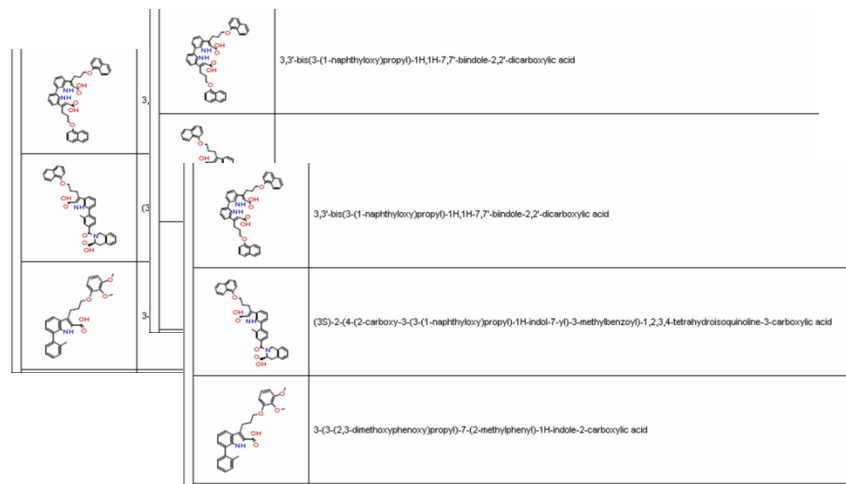
Extracted  
chemical  
names

# Mining for Chemical Knowledge – Use Case

Medicinal Chemist wants to synthesize competitor compound as tool compound for own project



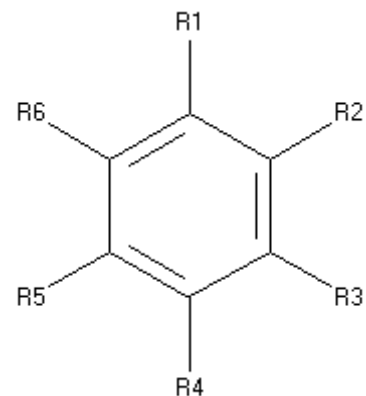
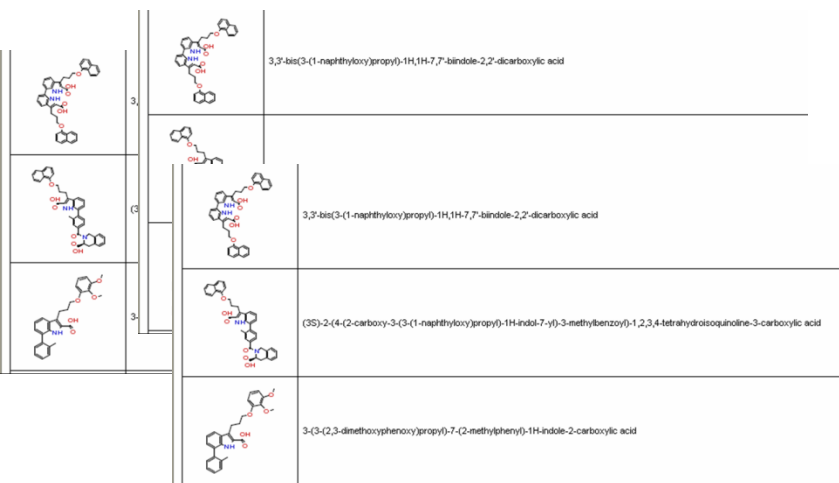
Extracted  
chemical  
names



Converted Chemical  
Structures

# Mining for Chemical Knowledge – Use Case

Medicinal Chemist wants to synthesize competitor compound as tool compound for own project



Converted Chemical Structures

Identification of core scaffold

# Mining for Chemical Knowledge – Use Case

Medicinal Chemist wants to synthesize competitor compound as tool compound for own project

**This enables the identification of compounds most representative for a competitor patent**

scaffold

Analysis of substitution patterns

# Mining for Chemical Knowledge

*Technologies from providers*

## Text entity recognition

- (a) Extractors (IUPAC names)
- TEMIS Chemical Entity Relationships Skill Cartridge
  - Accelrys Pipeline Pilot extractor (Notiora)
  - Fraunhofer (ProMiner Chemistry)
  - Chemaxon (chemicalize.org)
  - Oscar (Corbett, Murray-Rust, Teufel et al.)
  - SureChem
  - IBM Chemfrag Annotator
- (b) Converter  
(Names → connection table)
- CambridgeSoft name=struct
  - Openeye Lexichem
  - Chemaxon

## Image recognition

- OSRA (NIH)
- Clide Pro (Keymodule Ltd.)
- Fraunhofer chemoCR
- ChemReader

# Mining for Chemical Knowledge and other *TMS technologies*

Other annotators	Application / Tools
<p>(a) Entity extractors</p> <ul style="list-style-type: none"><li>- ALEx applied on<ul style="list-style-type: none"><li>Products</li><li>Gene names</li><li>Target names</li><li>Modes of action</li><li>Diseases</li><li>Companies</li><li>People</li></ul></li></ul> <p>(b) Pattern extractors applied on</p> <ul style="list-style-type: none"><li>Compound numbers</li><li>Gene/protein identifiers</li><li>Document numbers</li><li>Patent numbers</li></ul>	<ul style="list-style-type: none"><li>- Clipboard Analysis</li><li>- UltraLink</li><li>- Bio/Cheminformatics tools</li><li>- Datawarehouse of text mining annotations</li><li>- Data analysis tools</li></ul>

# Mining for Chemical Knowledge – Novartis Tools

The screenshot displays a software interface titled "Clipboard Analysis". On the left, a "Text Box" is visible. Below it, a list of chemical structures is shown, each with a checkbox and a label "Chemical Entity (154)". Three structures are checked. In the center, a large orange box contains the text "Clipboard Analysis". To the right, a large area of text represents patent text, with several lines highlighted in blue. Three callout boxes are overlaid on the interface: "Identified structures" points to the list of chemical structures; "Patent text" points to the highlighted text; and "View structure on MouseOver" points to a specific chemical structure that is highlighted in the patent text. At the bottom right, another callout box says "Export to other applications".

Identified structures

Patent text

View structure on MouseOver

Export to other applications

# Mining for Knowledge – Novartis Tools

## Input example: J Med Chem Paper



### Extraction

Orally Bioavailable Antagonists of **Inhibitor of Apoptosis** Proteins Based on an **Azabicyclooctane** Scaffold

- [Abstract](#)
- [Full Text HTML](#)
- [Hi-Res PDF\[668 KB\]](#)
- [PDF w/ Links\[262 KB\]](#)
- [Supporting Info](#)
- [Figures](#)
- [References](#)

Frederick Cohen<sup>1</sup>, Bruno Alicke<sup>1</sup>, Linda O. Elliott<sup>1</sup>, John A. Flygare<sup>1</sup>, Tatiana Goncharov<sup>1</sup>, Stephen F. Kotelas<sup>1</sup>, Matthew C. Franklin<sup>1</sup>, Stacy Frankovitz<sup>1</sup>, Jean-Philippe Stephan<sup>1</sup>, Vickie Teui<sup>1</sup>, Domagoj Vucic<sup>1</sup>, Harvey Wong<sup>2</sup> and Wayne J. Fairbrother<sup>1</sup>


Departments of Discovery Chemistry, Translational Oncology, Biochemical Pharmacology, Protein Engineering, Assay and Automation Technology, and Drug Metabolism and Pharmacokinetics, [Genentech, Inc.](#), 1 DNA Way, South San Francisco, California 94080

- ✓  Chemical Entity (80)
- ✓  Companies (13)
- ✓  Diseases (9)
- ✓  Gene Name (57)
- ✓  Modes of Action (2)
- ✓  Products (28)
- ✓  People (13)
- ✓  Targets (47)

Programmed cell death, or apoptosis, is a crucial mechanism for maintaining homeostasis and removal of damaged or malignant cells. This pathway is tightly controlled by a number of positive and negative regulatory elements. The **inhibitor of apoptosis (IAP)** is a

Abbreviations: **IAP**, inhibitor of apoptosis; **XIAP**, X-chromosome-linked inhibitor of apoptosis; **BIR**, baculoviral inhibitor of apoptosis repeat; **ML-IAP**, melanoma inhibitor of apoptosis; **SMAC**, second mitochondrial activator of **caspase**; **Teoc**, (trimethylsilyl)ethyl carbamate; **TFA**, trifluoroacetic acid; **TMS**, trimethylsilyl; **FDC**, *N*-(3-dimethylampropyl)-*N'*-ethylcarbamidate hydrochloride; **HATU**, 2-(7-aza-1-hydrobenzotriazol-1-yl)-1,3,3-tetramethyluronium hexafluorophosphate; **TASF**, *tris*(dimethylamino)sulfonium difluorotrimethylsilylate; **HMEC**, human mammary epithelial cells.

proteins negatively regulate this process through a variety of mechanisms including direct inhibition of effector **caspase** enzymes or modulation of TNF receptor-mediated signaling pathways.(1-6) Members of this family are up-regulated in various **cancers** and promote resistance to chemotherapy. Thus, inhibition of these proteins may be a new **therapeutic** mechanism for treating **cancer**.(7, 8)



Scheme 1. Synthesis of **Amide** Analogues **14a-c**

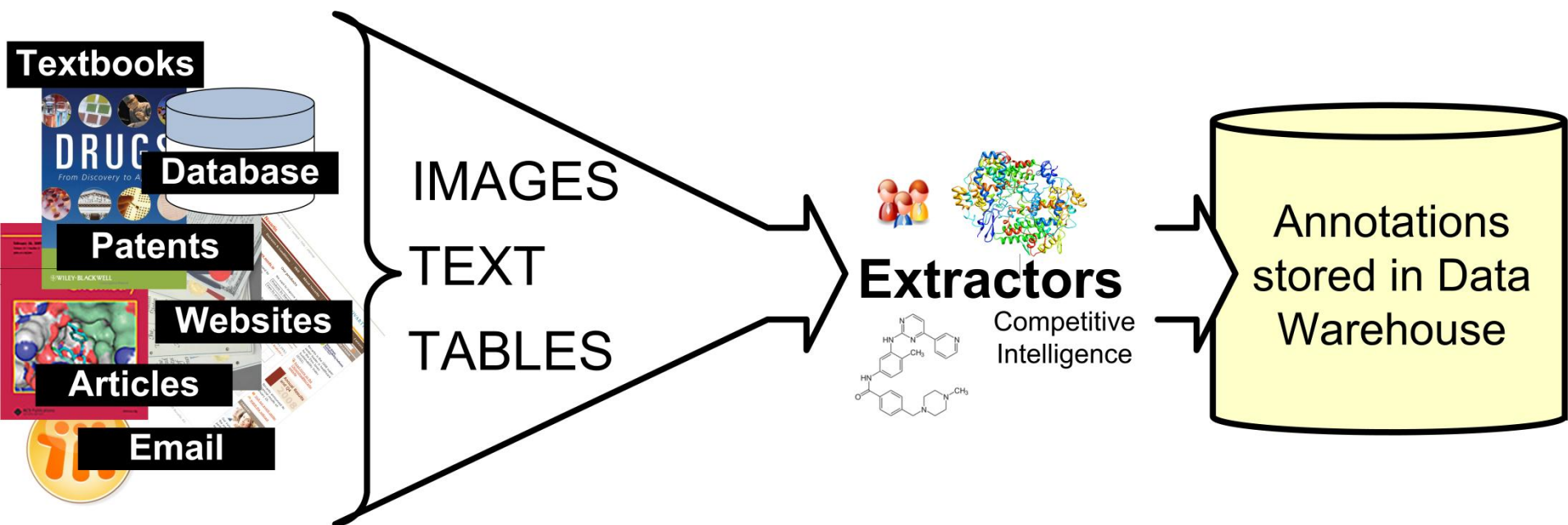
Each of the IAP proteins contains at least one BIR domain, some of which bind to the N-termini of their targets through interactions with a shallow cleft. Although the **peptide**-binding pocket of the BIR domains is shallow, several groups have reported peptidomimetics(9-12) or other small molecules(13-16) that bind to this pocket with high affinity, disrupt IAP-**caspase** or IAP-**SMAC** binding, and show in vivo efficacy in a mouse xenograft model of **cancer**.(17, 18)

In this communication we report our initial efforts toward developing a molecule with reduced **peptide** character that would possess the appropriate potency, permeability, and pharmacokinetic properties to be a useful **therapeutic** agent. To generate ideas for small molecule scaffolds that could potentially displace parts of the **peptide**, the software program CAVEAT was utilized.(19) This program takes as input the structure of a **peptide** or small molecule bound in the target receptor site. Selected bond vectors are searched against prebuilt databases of three-dimensional compounds to find ring systems that could connect to and maintain the vectors in their input spatial arrangement.

We used the 2.3 Å resolution crystal structure of **peptide** Ala-Val-Pro-2-diphenethylamine (1) bound to the **ML-IAP** BIR domain as the input structure (Figure 1A).(20) One vector was defined from the **proline nitrogen** atom to the **valine carbonyl carbon** atom, while a second vector was defined from the **proline carbonyl carbon** atom to the **diphenethylamine nitrogen** (Figure 2A). CAVEAT databases of minimized compounds from commercially available small molecules were constructed and searched. Ring systems that overlapped with atoms of the **ML-IAP** BIR domain were not considered in subsequent analyses.

# Integration chemical, biological knowledge

*Further TMS tools*



# Acknowledgements

---

## **Knowledge Engineers**

- Martin Romacker
- Pierre Parisot
- Josef Scheiber
- Nicolas Grandjean

## **Computer scientists**

- Alex Fromm
- Katia Vella
- Daniel Cronenberger
- Olivier Kreim

## **Cooperations**

- Steve Cleaver
- Greg Mccallister
- Jeremy Jenkins
- Dmitri Mikhailov
- Clayton Springer
- Naeem Yusuff
- Bharat Lagu

And many other people in different divisions of NIBR for their support