**Project Ungava**
Newton, Glen

NRC·CNRC

Canada Institute for Scientific and Technical Information

# Project Ungava

Glen Newton
Group Leader, CISTI Research
April 16 2007
CNI Spring Meeting Phoenix, AZ

NRC·CNRC
Canada Institute for Scientific and Technical Information

# Outline

- Context & Problem Space
- Long-term Strategic Objectives
- Short term objectives
- The catalog
- Search & navigation
- Clustering & visualization
- Semantic Text Mining supporting Domain-specific navigation & search

NRC·CNRC
Canada Institute for Scientific and Technical Information

# Context & Problem Space

- CISTI: Canada Institute for Scientific and Technical Information
- Canadian National Science Library
- Client groups: academic researchers, industry researchers, students, scientific / medical / engineering  innovators, etc.
- Also scientific publisher: NRC Research Press

NRC·CNRC
Canada Institute for Scientific and Technical Information

# CISTI's Strategic Plan

- *"Value proposition....An integrated "infostructure": electronic access to scientific information, using **intelligent search and analysis tools**"*
- *"Objectives...*
  - *1.4 Offer tools to **facilitate discovery and exploitation** of research*
  - *4.4 Conduct research in information science to advance knowledge and promote the adoption of new practices"*

NRC·CNRC
Canada Institute for Scientific and Technical Information

# CISTI Research

- Research group for CISTI
- Digital library research
- Three main research areas:
  - Text mining
  - Visualization
  - Agents

NRC·CNRC
Canada Institute for Scientific and Technical Information

# Long Term Strategic Objectives

- 5 – 10yr : knowledge not documents (but supported by documents)
- Discovery tools facilitating discovery of explicit and implicit actionable knowledge in literature, data, other
- Allow researchers to work in *their* language:
- *"Show me all the **compounds** which effect the glycolysis metabolic pathway by inhibiting kinases"*
- *"Show me all the **organisms** which are found in arctic soils which are collembolen predators and resistant to heavy metals"*
- *"Show me all the **compounds** which are non-metallic, have a melting point > 500C, tensile strength > 49.6 kg/mm/mm and electrical resistivity < 4600 Ω*m"*
- *“Show if soil mites are impacted by carbon tetrachloride in soils”*

NRC·CNRC
Canada Institute for Scientific and Technical Information

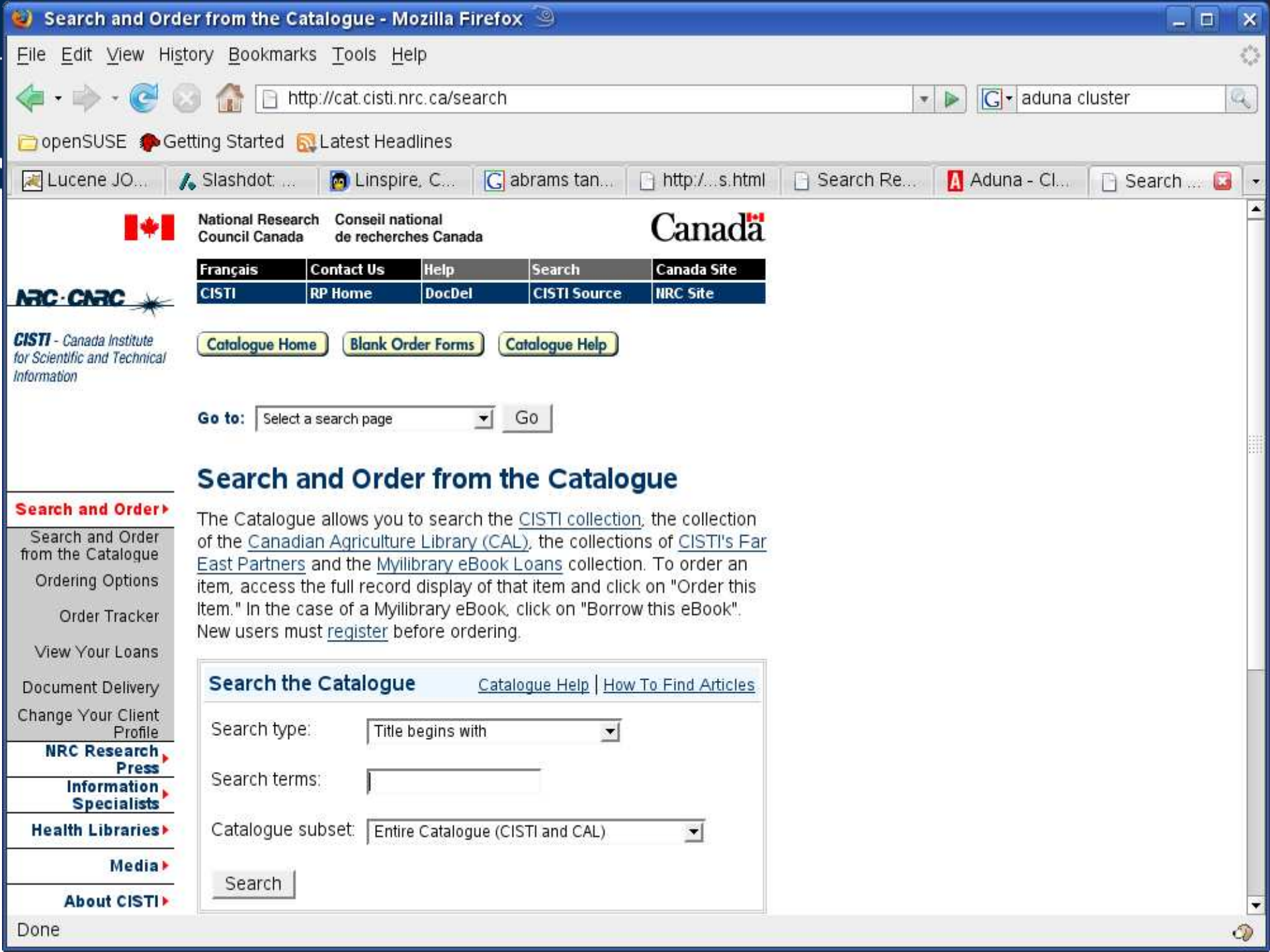
# Short Term Objectives

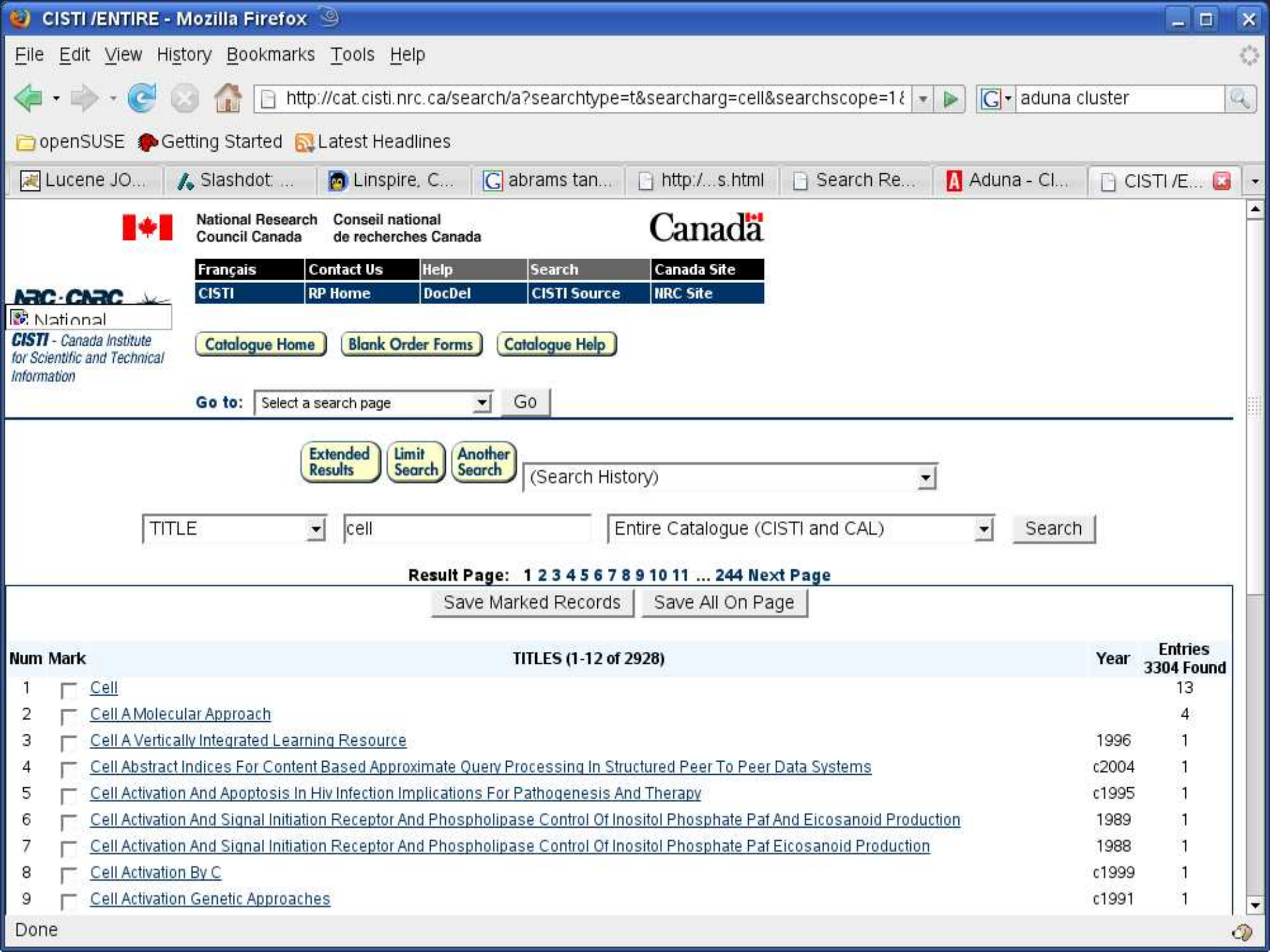
- Improve discovery (search, navigation & visualization) at CISTI
- Develop capacity, partnerships & prototypes in semantic web to extend discovery to offer richer semantic capabilities: domain specific search & navigation

NRC·CNRC
Canada Institute for Scientific and Technical Information

# Ungava

- Improving discovery in existing collections (JOS, Source, Catalog, etc)
- Leveraging best-of-breed Open Source systems like *Lucene, Carrot2, Aduna, Simile Project etc.*
- Scalable, flexible, usable, economical:
  - metadata & full-text search
  - Simple integrated navigational search
  - *Similar documents* functionality
  - Union search (single index for JOS, Source, Catalog, other)
  - Results clutsering & visualization, including **tag cloud**-like functionality from Web 2.0
- First target: low hanging fruit: the **catalog**

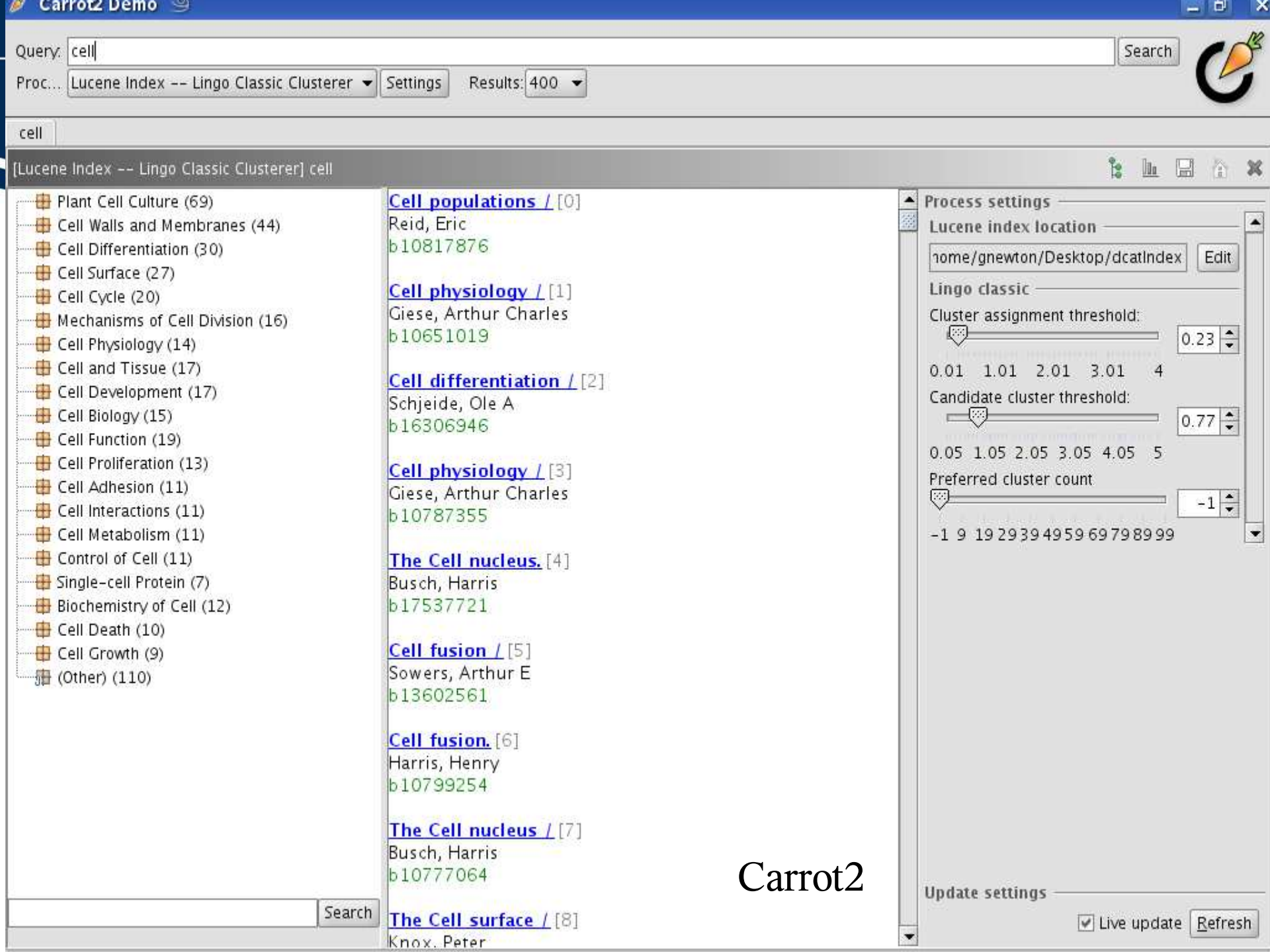

Search and Order from the Catalogue - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://cat.cisti.nrc.ca/search
aduna cluster
openSUSE Getting Started Latest Headlines
Lucene JO... Slashdot: ... Linspire, C... abrams tan... http:/...s.html Search Re... Aduna - Cl... Search ...
National Research Council Canada
Conseil national de recherches Canada
Canada
Français | Contact Us | Help | Search | Canada Site
CISTI | RP Home | DocDel | CISTI Source | NRC Site
NRC·CNRC
CISTI - Canada Institute for Scientific and Technical Information
Catalogue Home
Blank Order Forms
Catalogue Help
Go to:
Select a search page
Go
Search and Order from the Catalogue
The Catalogue allows you to search the CISTI collection, the collection of the Canadian Agriculture Library (CAL), the collections of CISTI's Far East Partners and the Myilibrary eBook Loans collection. To order an item, access the full record display of that item and click on "Order this Item." In the case of a Myilibrary eBook, click on "Borrow this eBook". New users must register before ordering.
Search and Order
Search and Order from the Catalogue
Ordering Options
Order Tracker
View Your Loans
Document Delivery
Change Your Client Profile
NRC Research Press
Information Specialists
Health Libraries
Media
About CISTI
Search the Catalogue
Catalogue Help | How To Find Articles
Search type:
Title begins with
Search terms:
Catalogue subset:
Entire Catalogue (CISTI and CAL)
Search
Done

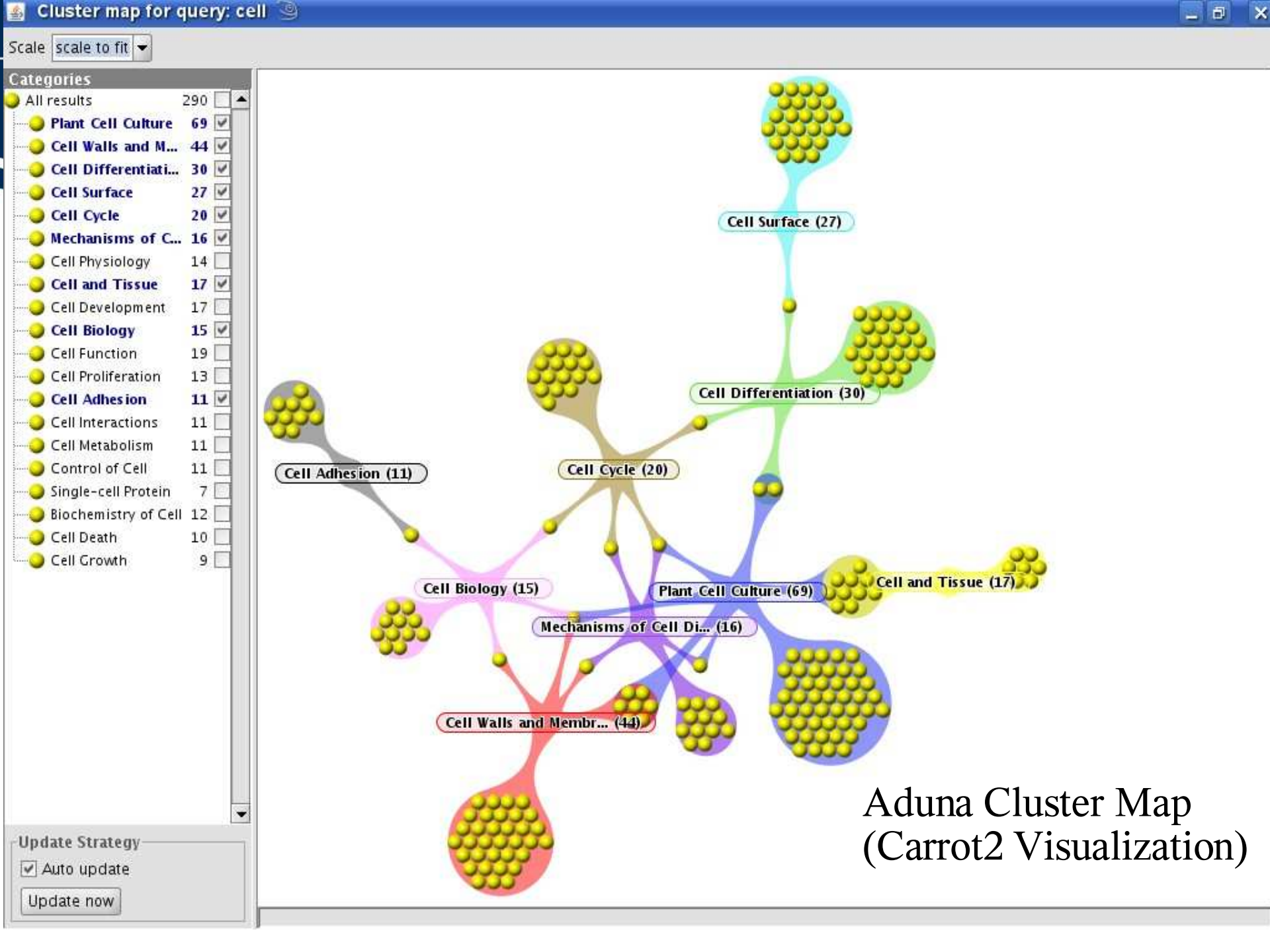

CISTI /ENTIRE - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://cat.cisti.nrc.ca/search/a?searchtype=t&searcharg=cell&searchscope=1
aduna cluster
openSUSE Getting Started Latest Headlines
Lucene JO... Slashdot: ... Linspire, C... abrams tan... http:/...s.html Search Re... Aduna - Cl... CISTI /E...
National Research Council Canada
Conseil national de recherches Canada
Canada
Français Contact Us Help Search Canada Site
CISTI RP Home DocDel CISTI Source NRC Site
NRC·CNRC
National
CISTI - Canada Institute for Scientific and Technical Information
Catalogue Home Blank Order Forms Catalogue Help
Go to: Select a search page Go
Extended Results Limit Search Another Search
(Search History)
TITLE cell Entire Catalogue (CISTI and CAL) Search
Result Page: 1 2 3 4 5 6 7 8 9 10 11 ... 244 Next Page
Save Marked Records Save All On Page
Num Mark TITLES (1-12 of 2928) Year Entries 3304 Found
1 Cell 13
2 Cell A Molecular Approach 4
3 Cell A Vertically Integrated Learning Resource 1996 1
4 Cell Abstract Indices For Content Based Approximate Query Processing In Structured Peer To Peer Data Systems c2004 1
5 Cell Activation And Apoptosis In Hiv Infection Implications For Pathogenesis And Therapy c1995 1
6 Cell Activation And Signal Initiation Receptor And Phospholipase Control Of Inositol Phosphate Paf And Eicosanoid Production 1989 1
7 Cell Activation And Signal Initiation Receptor And Phospholipase Control Of Inositol Phosphate Paf Eicosanoid Production 1988 1
8 Cell Activation By C c1999 1
9 Cell Activation Genetic Approaches c1991 1
Done

NRC·CNRC
Canada Institute for Scientific and Technical Information

# How?

- Get the content out of the (traditional) library catalog: XML export
- Index with Lucene
- Perform results clustering & visualization of results: Carrot2, Aduna, Simile


Carrot2 Demo
Query: cell
Search
Proc... Lucene Index -- Lingo Classic Clusterer
Settings
Results: 400
cell
[Lucene Index -- Lingo Classic Clusterer] cell
Plant Cell Culture (69)
Cell Walls and Membranes (44)
Cell Differentiation (30)
Cell Surface (27)
Cell Cycle (20)
Mechanisms of Cell Division (16)
Cell Physiology (14)
Cell and Tissue (17)
Cell Development (17)
Cell Biology (15)
Cell Function (19)
Cell Proliferation (13)
Cell Adhesion (11)
Cell Interactions (11)
Cell Metabolism (11)
Control of Cell (11)
Single-cell Protein (7)
Biochemistry of Cell (12)
Cell Death (10)
Cell Growth (9)
(Other) (110)
Search
Cell populations / [0]
Reid, Eric
b10817876
Cell physiology / [1]
Giese, Arthur Charles
b10651019
Cell differentiation / [2]
Schjeide, Ole A
b16306946
Cell physiology / [3]
Giese, Arthur Charles
b10787355
The Cell nucleus. [4]
Busch, Harris
b17537721
Cell fusion / [5]
Sowers, Arthur E
b13602561
Cell fusion. [6]
Harris, Henry
b10799254
The Cell nucleus / [7]
Busch, Harris
b10777064
The Cell surface / [8]
Knox, Peter
Process settings
Lucene index location
nome/gnewton/Desktop/dcatIndex
Edit
Lingo classic
Cluster assignment threshold:
0.23
0.01 1.01 2.01 3.01 4
Candidate cluster threshold:
0.77
0.05 1.05 2.05 3.05 4.05 5
Preferred cluster count
-1
-1 9 19 29 39 49 59 69 79 89 99
Update settings
Live update
Refresh
Carrot2


Cluster map for query: cell
Scale
scale to fit
Categories
All results 290
Plant Cell Culture 69
Cell Walls and M... 44
Cell Differentiati... 30
Cell Surface 27
Cell Cycle 20
Mechanisms of C... 16
Cell Physiology 14
Cell and Tissue 17
Cell Development 17
Cell Biology 15
Cell Function 19
Cell Proliferation 13
Cell Adhesion 11
Cell Interactions 11
Cell Metabolism 11
Control of Cell 11
Single-cell Protein 7
Biochemistry of Cell 12
Cell Death 10
Cell Growth 9
Update Strategy
Auto update
Update now
Cell Surface (27)
Cell Differentiation (30)
Cell Adhesion (11)
Cell Cycle (20)
Cell Biology (15)
Plant Cell Culture (69)
Cell and Tissue (17)
Mechanisms of Cell Di... (16)
Cell Walls and Membr... (44)
Aduna Cluster Map
(Carrot2 Visualization)

NRC·CNRC
Canada Institute for Scientific and Technical Information

# Search Box Fallacy

- Assumption that the single (or multiple entry) search box solves all use cases
- Solves some, probably many use cases
- Alternate views, modalities (like visualization, etc)
- Important other ones: personal collections:
  - "I found these 3 documents that are just what I was looking for (elsewhere). Show me all the documents in *your* collection that are like this"
  - "I found these 4 web pages…
- Similarities to user's content or sub-collections

NRC·CNRC
Canada Institute for Scientific and Technical Information

# KDSD-SITMSL Sub-project

- Knowledge Discovery Supported Domain-Specific and Interdisciplinary Text Mining of Scientific Literature
- Use taxonomies/classifications/thesaurii etc from research domain to offer enhanced search & navigation to researchers from that domain (Domain-specific faceted search & browse)
- Support interdisciplinary discovery
- Text mine out entities & create search & navigation taxonomies
- Initial domains: Biological taxonomy (classification), minerological classification, Gene ontology, geological epoch, MeSH terms

NRC·CNRC
Canada Institute for Scientific and Technical Information

# KDSD-SITMSL Sub-project

- Knowledge Discovery Supported Domain-Specific and Interdisciplinary Text Mining of Scientific Literature
- Use taxonomies/classifications/thesaurii etc from research domain to offer enhanced search & navigation to researchers from that domain (Domain-specific faceted search & browse)
- Support interdisciplinary discovery
- Text mine out entities & create search & navigation taxonomies
- Initial domains: Biological taxonomy (classification), minerological classification, Gene ontology, geological epoch, MeSH terms

NRC·CNRC
Canada Institute for Scientific and Technical Information

# Next steps

- Continue with building base
- NSERC grant application with U of Ottawa & Carleton University
- Examine new Semantic Web inference engines
- Look at alternate full-text engines like Terrier, etc.
- Migration from CISTI Research to CISTI production of some ideas

NRC·CNRC
Canada Institute for Scientific and Technical Information

# Acknowledgements


- Jeff Demaine, Greg Kresko, Dr. Andre Vellino: CISTI Research
- Dr. Val Behan-Pelletier, Dr. Guy Baillergeon (ITIS database): Agriculture Canada
- Dr. Diana Inkpen, Unicersity of Ottawa
- Dr. Michel Dumontier, Carleton Universityv

NRC·CNRC
Canada Institute for Scientific and Technical Information

# Questions?

- Glen Newton glen.newton@nrc-cnrc.gc.ca

NRC·CNRC

Canada Institute for Scientific and Technical Information

# Science at work for Canada

National Research Council Canada

Conseil national de recherches Canada

Canada