ChemXSeer: CyberInfrastructure for Chemical Kinetics

Prasenjit Mitra
College of Information Sciences and Technology,
Department of Computer Science and Engineering

In collaboration with: C. Lee Giles, J. Bandstra, K. Mueller, J. Kubicki, S. Brantley, B. Brouwer, S. Nangia, B. Garrison, B. Sun, Y. Liu, Q. Tan, A.R. Jaiswal, Juan Fernandez, L. Bolleli, X. Lu, …

Motivation

- Data sharing
  - Data repository
  - Querying and finding data efficiently
  - Analysis tools
  - Preservation and archival
- Data extraction
  - From tables
  - From figures

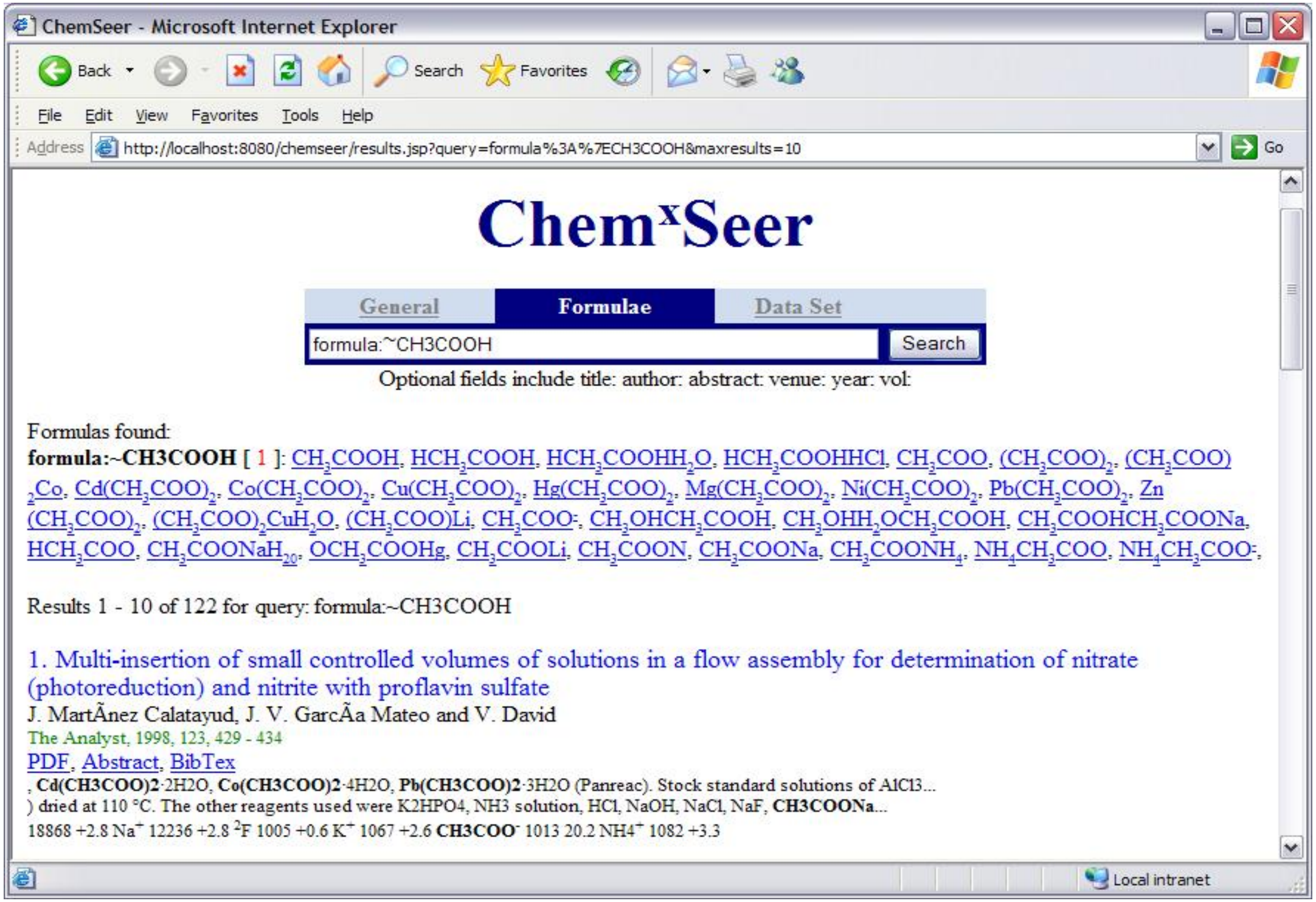
Vertical Search Engine

- Domain-specific search engine
  - Entity Extraction
  - Indexing
  - Ranking
    - Similarity & Relevance
- ChemXSeer
  - Chemical Formulae and Names
  - Different query semantics
    - Fuzzy search, similarity search
- What are the first-class entities important for your grand challenge?
  - Existing work on diseases, proteins
  - Genes, Enzymes, ...

Searching Documents

- Efficient search tools
  - Chemical entity search
    - Formula
    - Name
    - Structure

Formula Search -Interface



http://localhost:8080/chemseer/results.jsp?query=formula%3A%7ECH3COOH&maxresults=10

Chem

formula:~CH3COOH

Optional fields include title: author: abstract: venue: year: vol:

Formulas found:

formula:~CH3COOH [1]: CH_3COOH, HCH_3COOH, HCH_3COOHH_2O, $HCH_3COOHHCl$, CH_3COO, $(CH_3COO)_2$, $(CH_3COO)_2Co$, $Cd(CH_3COO)_2$, $Co(CH_3COO)_2$, $Cu(CH_3COO)_2$, $Hg(CH_3COO)_2$, $Mg(CH_3COO)_2$, $Ni(CH_3COO)_2$, $Pb(CH_3COO)_2$, $Zn(CH_3COO)_2$, $(CH_3COO)_2CuH_2O$, $(CH_3COO)Li$, CH_3COO^-, CH_3OHCH_3COOH, $CH_3OHH_2OCH_3COOH$, $CH_3COOHCH_3COONa$, HCH_3COO, $CH_3COONaH_{20}$, OCH_3COOHg, CH_3COOLi, CH_3COON, CH_3COONa, CH_3COONH_4, NH_4CH_3COO, $NH_4CH_3COO^-$,

Results 1 - 10 of 122 for query: formula:~CH3COOH

1. Multi-insertion of small controlled volumes of solutions in a flow assembly for determination of nitrate (photoreduction) and nitrite with proflavin sulfate

J. MartÃnez Calatayud, J. V. GarcÃa Mateo and V. David

The Analyst, 1998, 123, 429 - 434

PDF, Abstract, BibTex

, **Cd(CH3COO)2**·2H2O, **Co(CH3COO)2**·4H2O, **Pb(CH3COO)2**·3H2O (Panreac). Stock standard solutions of AlCl3...
) dried at 110 °C. The other reagents used were K2HPO4, NH3 solution, HCl, NaOH, NaCl, NaF, **CH3COONa**...
18868 +2.8 Na^+ 12236 +2.8 ^2F 1005 +0.6 K^+ 1067 +2.6 **CH3COO**$^-$ 1013 20.2 $NH4^+$ 1082 +3.3

Chemical Entity Search

- Search engines do not understand chemical formulae, chemical names
- No fuzzy search capabilities
  - With functional groups, e.g., -OH
- Automatic segmentation of chemical names and indexing
- Structure search algorithms need improvement

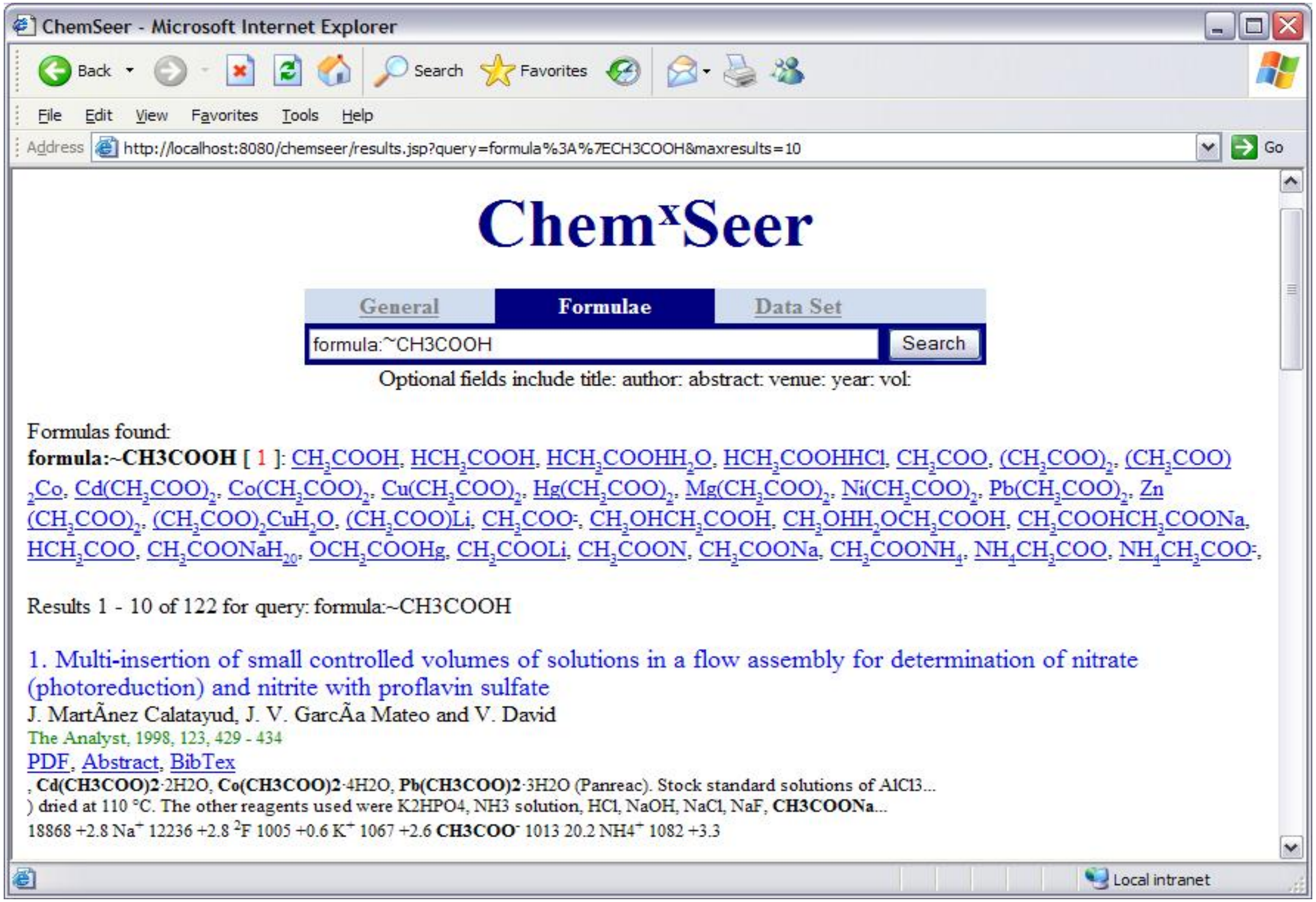
Formula Search -Query Models

- Substructure search
  - Search for formulae that may have a substructure
  - E.g. -COOH matches CH3COOH (exact match: high score), HOOCCH3 (reverse match: medium score), and CH3CHO2 (parsed match: low score).
- Similarity search
  - Search for formulae with a similar structure of the query formula. Feature-based approach using partial formulae matching.
  - E.g. ~CH3COOH matches CH3COOH, (CH3COO)2Co, CH3COO$^-$, etc.

Formula Search -Query Models

- Conjunctive search of the four types of formula searches
  - E.g. [*C2H4-6 -COOH] matches CH3COOH, not C2H4O or CH3CH2COOH.
- Document query rewriting
  - E.g. document query water formula:=CH4 is rewritten to water (CH4 OR H4C OR CD4), if formula search of =CH4 matches CH4, H4C and CD4.

Experiments -Interface



http://localhost:8080/chemseer/results.jsp?query=formula%3A%7ECH3COOH&maxresults=10

Chem^{x}Seer

formula:~CH3COOH

Optional fields include title: author: abstract: venue: year: vol:

Formulas found:
formula:~CH3COOH [1]: CH_3COOH, HCH_3COOH, HCH_3COOHH_2O, $HCH_3COOHHCl$, CH_3COO, $(CH_3COO)_2$, $(CH_3COO)_2Co$, $Cd(CH_3COO)_2$, $Co(CH_3COO)_2$, $Cu(CH_3COO)_2$, $Hg(CH_3COO)_2$, $Mg(CH_3COO)_2$, $Ni(CH_3COO)_2$, $Pb(CH_3COO)_2$, $Zn(CH_3COO)_2$, $(CH_3COO)_2CuH_2O$, $(CH_3COO)Li$, CH_3COO^-, CH_3OHCH_3COOH, $CH_3OHH_2OCH_3COOH$, $CH_3COOHCH_3COONa$, HCH_3COO, $CH_3COONaH_{20}$, OCH_3COOHg, CH_3COOLi, CH_3COON, CH_3COONa, CH_3COONH_4, NH_4CH_3COO, $NH_4CH_3COO^-$,

Results 1 - 10 of 122 for query: formula:~CH3COOH

1. Multi-insertion of small controlled volumes of solutions in a flow assembly for determination of nitrate (photoreduction) and nitrite with proflavin sulfate
J. MartÃnez Calatayud, J. V. GarcÃa Mateo and V. David
The Analyst, 1998, 123, 429 - 434
PDF, Abstract, BibTex
, **Cd(CH3COO)2**·2H2O, **Co(CH3COO)2**·4H2O, **Pb(CH3COO)2**·3H2O (Panreac). Stock standard solutions of AlCl3...
) dried at 110 °C. The other reagents used were K2HPO4, NH3 solution, HCl, NaOH, NaCl, NaF, **CH3COONa**...
18868 +2.8 Na^+ 12236 +2.8 ^2F 1005 +0.6 K^+ 1067 +2.6 **CH3COO**$^-$ 1013 20.2 $NH4^+$ 1082 +3.3

Data Repository

- Store and publish data
  - Gaussian
  - CHARMM
  - Excel data
    - Soil profiles
    - Dissolution rates
  - Spectroscopy data

Functionality

- Store in databases
  - Fast access
  - Structured access
    - Query conditions, e.g., 277 < temp < 302
    - Combine information from multiple tables
  - Query against multiple formats
    - Mediated architecture

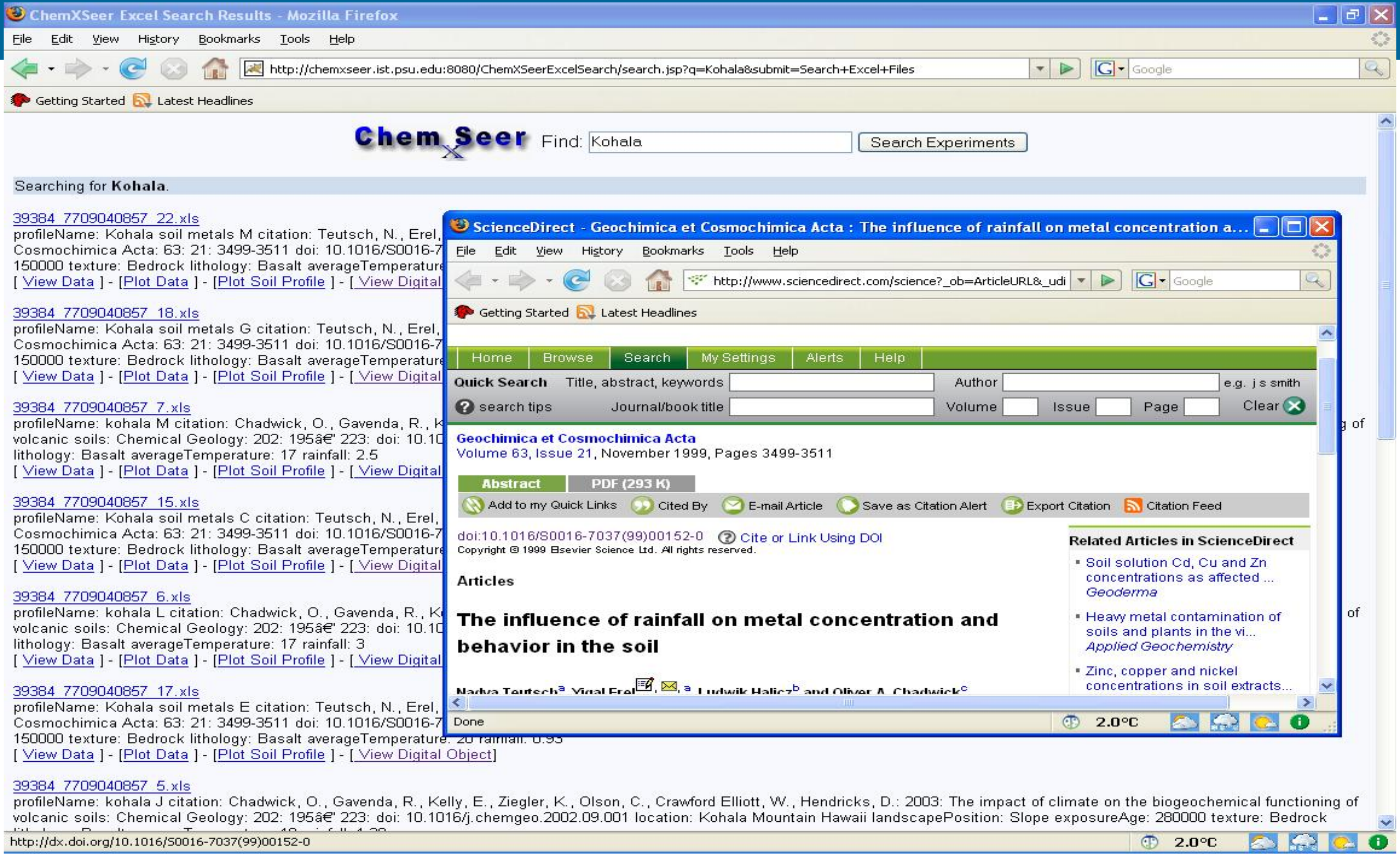
Challenges

- Automatic processing needs fixed data formats
- Ideally one column name
  - Distinct column names across one table
  - Or at least a fixed number of lines for the table header before the data starts
- Fully empty columns and rows are confusing
- Use different datasheets for different tables
- Embedding formulae in the dataset is confusing

Metadata

- Data describing data
- You need to tell the computer what the columns mean
  - What data is contained in the table?
  - What is the relationship of the different columns and between different sheets of data in Excel?
  - Where was the data obtained?
  - Any other information that may be useful for search and querying later on
- Metadata needs to be in a fixed, specified format
  - Use a form

Data Search - Interface



ChemXSeer Excel Search Results - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://chemxseer.ist.psu.edu:8080/ChemXSeerExcelSearch/search.jsp?q=Kohala&submit=Search+Excel+Files
Getting Started Latest Headlines
ChemXSeer Find: Kohala
Search Experiments

Searching for **Kohala**.

39384_7709040857_22.xls
profileName: Kohala soil metals M citation: Teutsch, N., Erel,
Cosmochimica Acta: 63: 21: 3499-3511 doi: 10.1016/S0016-7
150000 texture: Bedrock lithology: Basalt averageTemperature
[View Data] - [Plot Data] - [Plot Soil Profile] - [View Digital

39384_7709040857_18.xls
profileName: Kohala soil metals G citation: Teutsch, N., Erel,
Cosmochimica Acta: 63: 21: 3499-3511 doi: 10.1016/S0016-7
150000 texture: Bedrock lithology: Basalt averageTemperature
[View Data] - [Plot Data] - [Plot Soil Profile] - [View Digital

39384_7709040857_7.xls
profileName: kohala M citation: Chadwick, O., Gavenda, R., K ... g of
volcanic soils: Chemical Geology: 202: 195â€" 223: doi: 10.10
lithology: Basalt averageTemperature: 17 rainfall: 2.5
[View Data] - [Plot Data] - [Plot Soil Profile] - [View Digital

39384_7709040857_15.xls
profileName: Kohala soil metals C citation: Teutsch, N., Erel,
Cosmochimica Acta: 63: 21: 3499-3511 doi: 10.1016/S0016-7
150000 texture: Bedrock lithology: Basalt averageTemperature
[View Data] - [Plot Data] - [Plot Soil Profile] - [View Digital

39384_7709040857_6.xls
profileName: kohala L citation: Chadwick, O., Gavenda, R., K ... of
volcanic soils: Chemical Geology: 202: 195â€" 223: doi: 10.10
lithology: Basalt averageTemperature: 17 rainfall: 3
[View Data] - [Plot Data] - [Plot Soil Profile] - [View Digital

39384_7709040857_17.xls
profileName: Kohala soil metals E citation: Teutsch, N., Erel,
Cosmochimica Acta: 63: 21: 3499-3511 doi: 10.1016/S0016-7
150000 texture: Bedrock lithology: Basalt averageTemperature: 20 rainfall: 0.95
[View Data] - [Plot Data] - [Plot Soil Profile] - [View Digital Object]

39384_7709040857_5.xls
profileName: kohala J citation: Chadwick, O., Gavenda, R., Kelly, E., Ziegler, K., Olson, C., Crawford Elliott, W., Hendricks, D.: 2003: The impact of climate on the biogeochemical functioning of volcanic soils: Chemical Geology: 202: 195â€" 223: doi: 10.1016/j.chemgeo.2002.09.001 location: Kohala Mountain Hawaii landscapePosition: Slope exposureAge: 280000 texture: Bedrock

ScienceDirect - Geochimica et Cosmochimica Acta : The influence of rainfall on metal concentration a...
File Edit View History Bookmarks Tools Help
http://www.sciencedirect.com/science?_ob=ArticleURL&_udi
Getting Started Latest Headlines
Home Browse Search My Settings Alerts Help
Quick Search Title, abstract, keywords Author e.g. j s smith
search tips Journal/book title Volume Issue Page Clear
Abstract PDF (293 K)
Add to my Quick Links Cited By E-mail Article Save as Citation Alert Export Citation Citation Feed
Cite or Link Using DOI
Related Articles in ScienceDirect
Done 2.0°C

Geochimica et Cosmochimica Acta
Volume 63, Issue 21, November 1999, Pages 3499-3511

doi:10.1016/S0016-7037(99)00152-0
Copyright © 1999 Elsevier Science Ltd. All rights reserved.

Articles

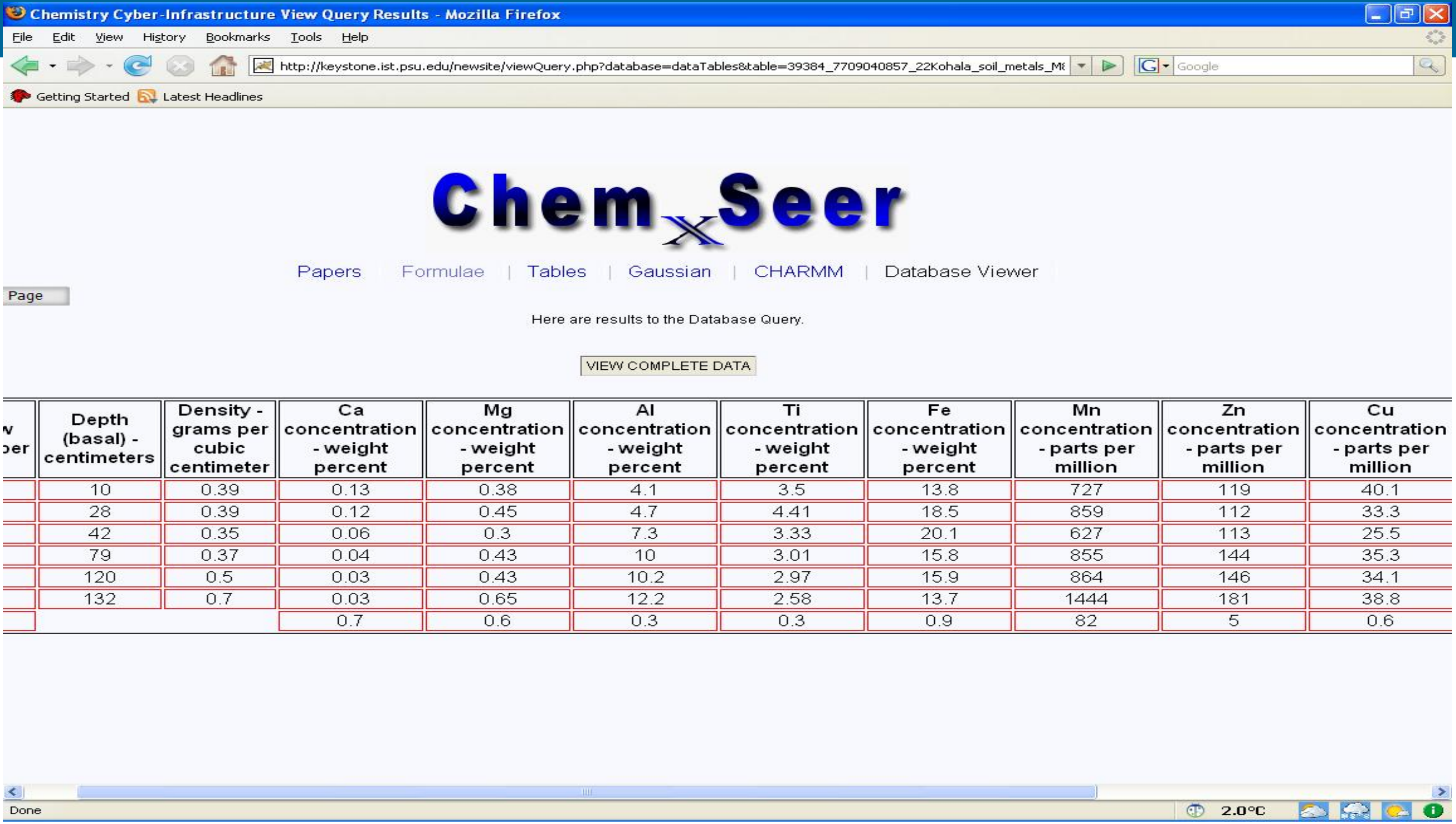
The influence of rainfall on metal concentration and behavior in the soil

Nadya Teutsch[a] Yigal Erel[a] Ludwik Halicz[b] and Oliver A. Chadwick[c]

- Soil solution Cd, Cu and Zn concentrations as affected ... *Geoderma*
- Heavy metal contamination of soils and plants in the vi... *Applied Geochemistry*
- Zinc, copper and nickel concentrations in soil extracts...

http://dx.doi.org/10.1016/S0016-7037(99)00152-0

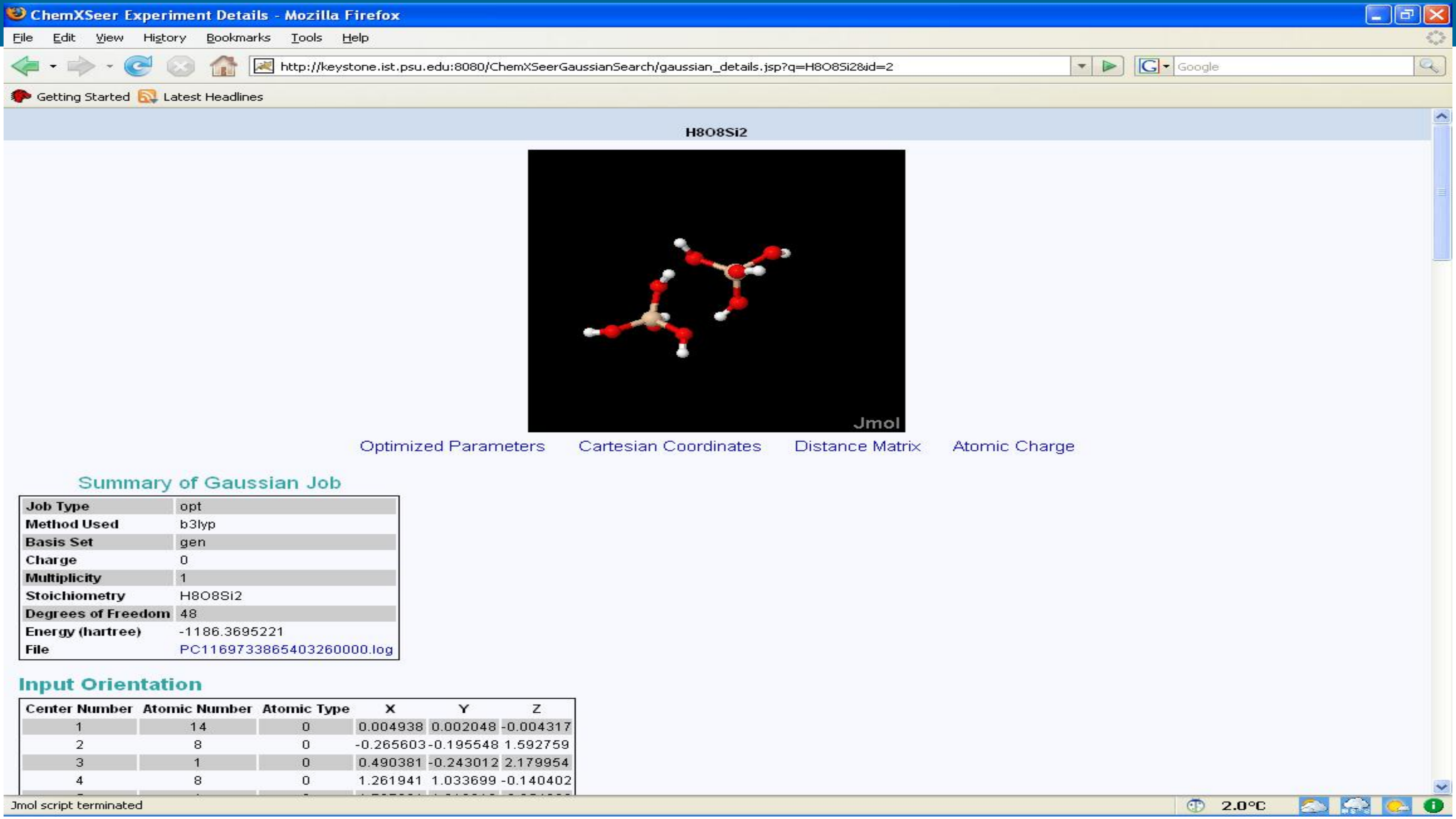
Data Search - Report


Chemistry Cyber-Infrastructure View Query Results - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://keystone.ist.psu.edu/newsite/viewQuery.php?database=dataTables&table=39384_7709040857_22Kohala_soil_metals_Mt
Google
Getting Started
Latest Headlines
ChemxSeer
Papers
Formulae | Tables | Gaussian | CHARMM | Database Viewer
Page
VIEW COMPLETE DATA
Done
2.0°C

Here are results to the Database Query.

w per	Depth (basal) - centimeters	Density - grams per cubic centimeter	Ca concentration - weight percent	Mg concentration - weight percent	Al concentration - weight percent	Ti concentration - weight percent	Fe concentration - weight percent	Mn concentration - parts per million	Zn concentration - parts per million	Cu concentration - parts per million
	10	0.39	0.13	0.38	4.1	3.5	13.8	727	119	40.1
	28	0.39	0.12	0.45	4.7	4.41	18.5	859	112	33.3
	42	0.35	0.06	0.3	7.3	3.33	20.1	627	113	25.5
	79	0.37	0.04	0.43	10	3.01	15.8	855	144	35.3
	120	0.5	0.03	0.43	10.2	2.97	15.9	864	146	34.1
	132	0.7	0.03	0.65	12.2	2.58	13.7	1444	181	38.8
			0.7	0.6	0.3	0.3	0.9	82	5	0.6

An Example


ChemXSeer Experiment Details - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Google
Getting Started Latest Headlines
Jmol
Jmol script terminated
2.0°C

http://keystone.ist.psu.edu:8080/ChemXSeerGaussianSearch/gaussian_details.jsp?q=H8O8Si2&id=2

H8O8Si2

Optimized Parameters Cartesian Coordinates Distance Matrix Atomic Charge

Summary of Gaussian Job

Job Type	opt
Method Used	b3lyp
Basis Set	gen
Charge	0
Multiplicity	1
Stoichiometry	H8O8Si2
Degrees of Freedom	48
Energy (hartree)	-1186.3695221
File	PC1169733865403260000.log

Input Orientation

Center Number	Atomic Number	Atomic Type	X	Y	Z
1	14	0	0.004938	0.002048	-0.004317
2	8	0	-0.265603	-0.195548	1.592759
3	1	0	0.490381	-0.243012	2.179954
4	8	0	1.261941	1.033699	-0.140402

Data Search - Plot

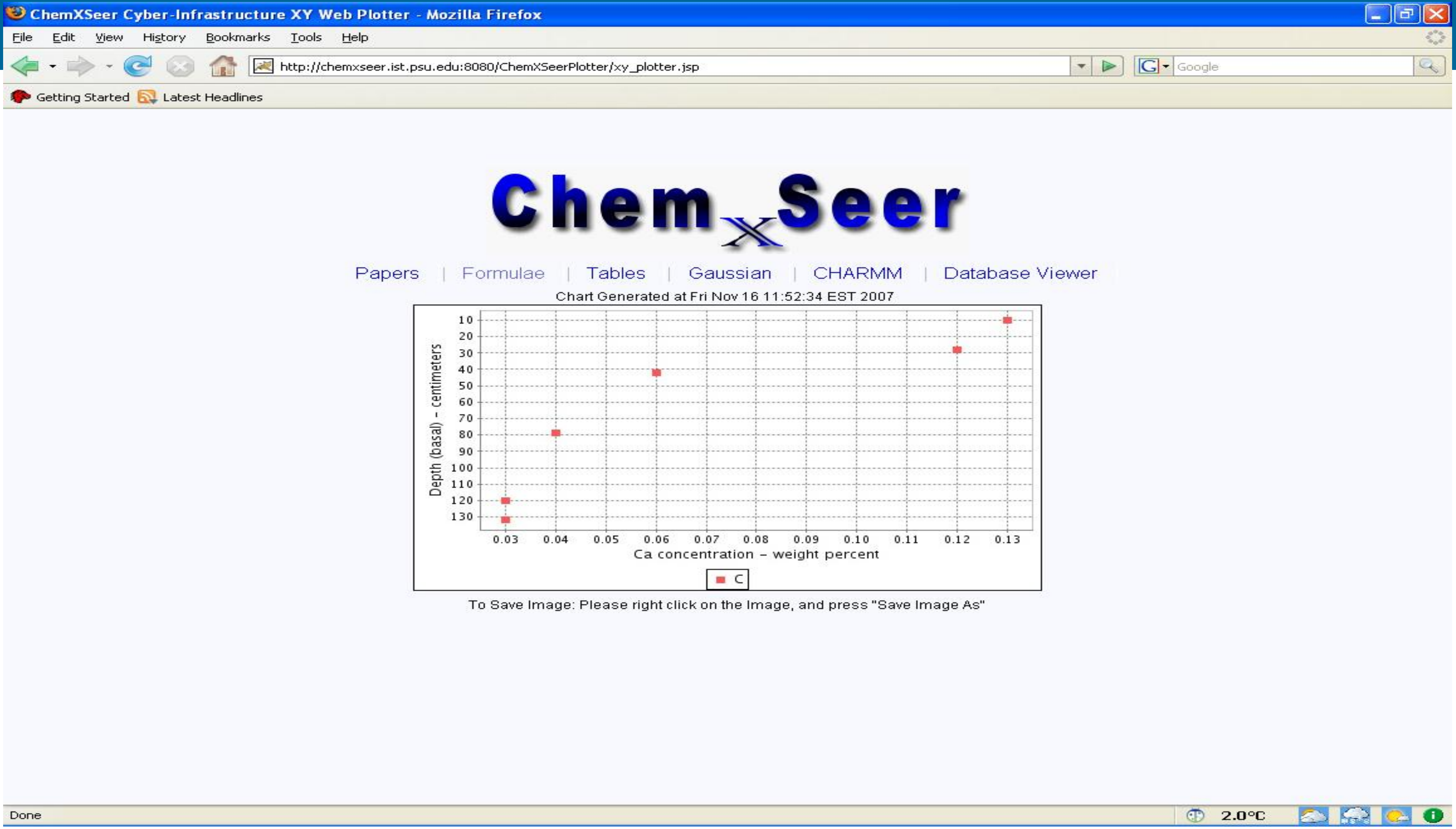

ChemXSeer Cyber-Infrastructure XY Web Plotter - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://chemxseer.ist.psu.edu:8080/ChemXSeerPlotter/xy_plotter.jsp
Google
Getting Started
Latest Headlines
ChemXSeer
Papers | Formulae | Tables | Gaussian | CHARMM | Database Viewer
Depth (basal) - centimeters
10
20
30
40
50
60
70
80
90
100
110
120
130
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.10
0.11
0.12
0.13
Ca concentration - weight percent
C
Done
2.0°C

Chart Generated at Fri Nov 16 11:52:34 EST 2007

To Save Image: Please right click on the Image, and press "Save Image As"

Data Search

- All terms used in the datasheets will be indexed
- Search using those terms will retrieve results
- Then, you can go into an individual sheet and write your queries
- Reports
  - Massage the data & plot
- Command-line interface
  - In SQL

Data Security

- Login
- Mark data as private or public
- Creation of groups
  - Share data only within your group

Linking Data and Documents

- Create Digital Object
  - Datasets
  - Documents
  - Supplementary notes
  - Supplementary data
  - Other information
- Data sets use document-identifier
- Document metadata contains pointer to data

Figure Search

- Figures are important sources of data and information
- Search for figures by keywords
- Index
  - Captions
  - References to the images
  - Legend/text in images (?)
  - Document-level metadata
    - Title, author, venue, year, publisher

Data Extraction from Figures

- Axes detection
- Legend detection
- Data point identification
  - Shape detection
  - Resolving overlapping points
- Tic detection
- Label detection
- Identifying the continuation of lines at points of intersection

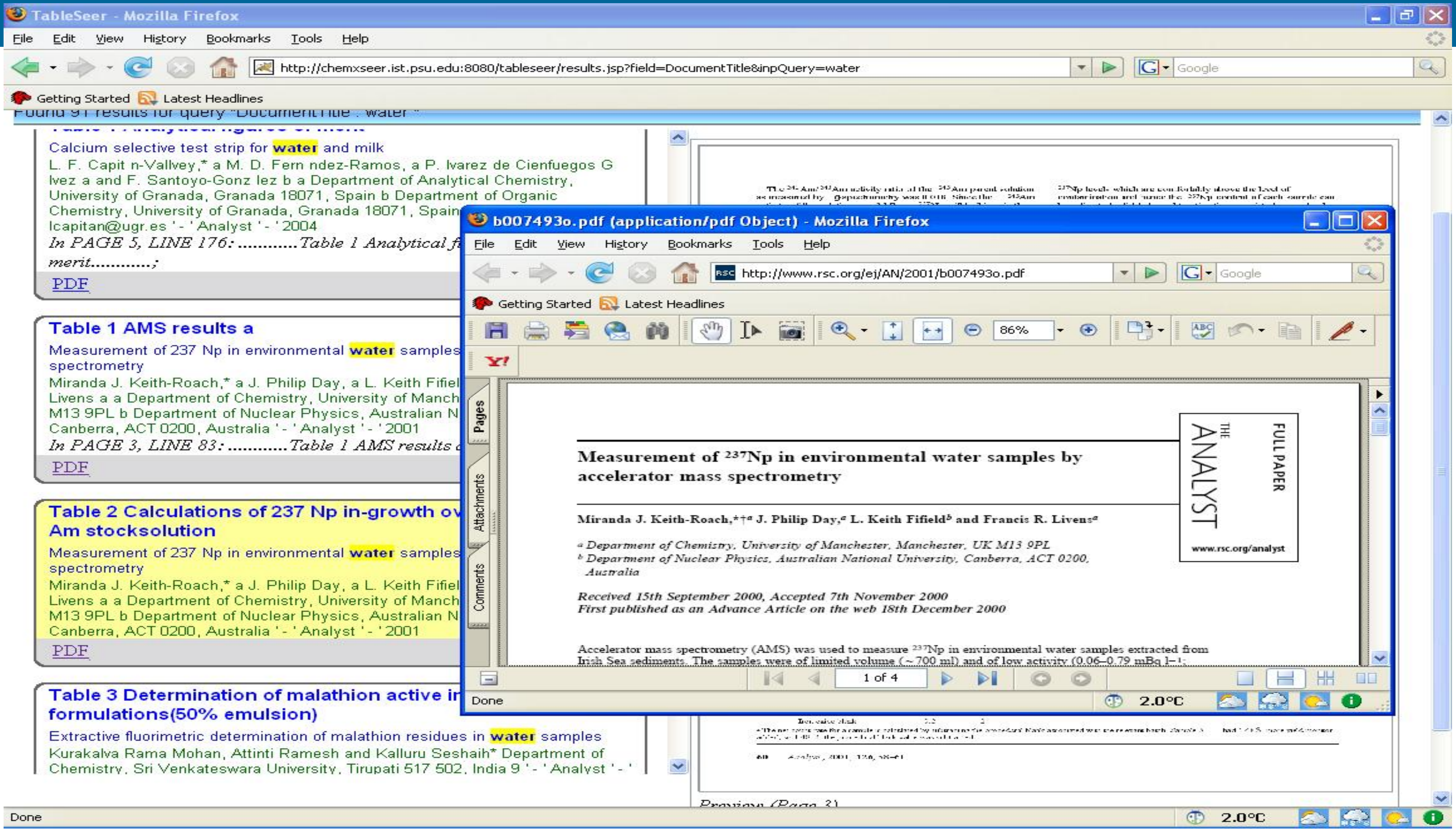
Challenges

- Optical Character Recognition (OCR)
  - Small letters
  - Sparse letters
- Overlapping shapes
- Axes Scale
  - Log scale
  - Linear scale
  - Disconnected
- Inversion of dependent and independent variables

TableSeer

- Search for tables with interesting data
- Extract data into a database
- Plot the data and support analysis
- Currently:
  - Table Search based on keywords
- Future:
  - Data extracted inserted into a database

TableSeer - Interface



TableSeer - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Getting Started Latest Headlines

http://chemxseer.ist.psu.edu:8080/tableseer/results.jsp?field=DocumentTitle&inpQuery=water

Found 91 results for query "DocumentTitle : water "

Calcium selective test strip for **water** and milk
L. F. Capit n-Vallvey,* a M. D. Fern ndez-Ramos, a P. lvarez de Cienfuegos G lvez a and F. Santoyo-Gonz lez b a Department of Analytical Chemistry, University of Granada, Granada 18071, Spain b Department of Organic Chemistry, University of Granada, Granada 18071, Spain
lcapitan@ugr.es '-' Analyst '-' 2004
In PAGE 5, LINE 176: ....Table 1 Analytical f merit............;
PDF

Table 1 AMS results a
Measurement of 237 Np in environmental **water** samples spectrometry
Miranda J. Keith-Roach,* a J. Philip Day, a L. Keith Fifiel Livens a a Department of Chemistry, University of Manch M13 9PL b Department of Nuclear Physics, Australian N Canberra, ACT 0200, Australia '-' Analyst '-' 2001
In PAGE 3, LINE 83: ....Table 1 AMS results a
PDF

Table 2 Calculations of 237 Np in-growth ov Am stocksolution
Measurement of 237 Np in environmental **water** samples spectrometry
Miranda J. Keith-Roach,* a J. Philip Day, a L. Keith Fifiel Livens a a Department of Chemistry, University of Manch M13 9PL b Department of Nuclear Physics, Australian N Canberra, ACT 0200, Australia '-' Analyst '-' 2001
PDF

Table 3 Determination of malathion active in formulations(50% emulsion)
Extractive fluorimetric determination of malathion residues in **water** samples
Kurakalva Rama Mohan, Attinti Ramesh and Kalluru Seshaih* Department of Chemistry, Sri Venkateswara University, Tirupati 517 502, India 9 '-' Analyst '-'

b007493o.pdf (application/pdf Object) - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Getting Started Latest Headlines
86%
Pages
Attachments
Comments
1 of 4
Done
2.0°C
Done
2.0°C

http://www.rsc.org/ej/AN/2001/b007493o.pdf

Measurement of ^{237}Np in environmental water samples by accelerator mass spectrometry

Miranda J. Keith-Roach,*†[a] J. Philip Day,[a] L. Keith Fifield[b] and Francis R. Livens[a]

[a] *Department of Chemistry, University of Manchester, Manchester, UK M13 9PL*
[b] *Department of Nuclear Physics, Australian National University, Canberra, ACT 0200, Australia*

Received 15th September 2000, Accepted 7th November 2000
First published as an Advance Article on the web 18th December 2000

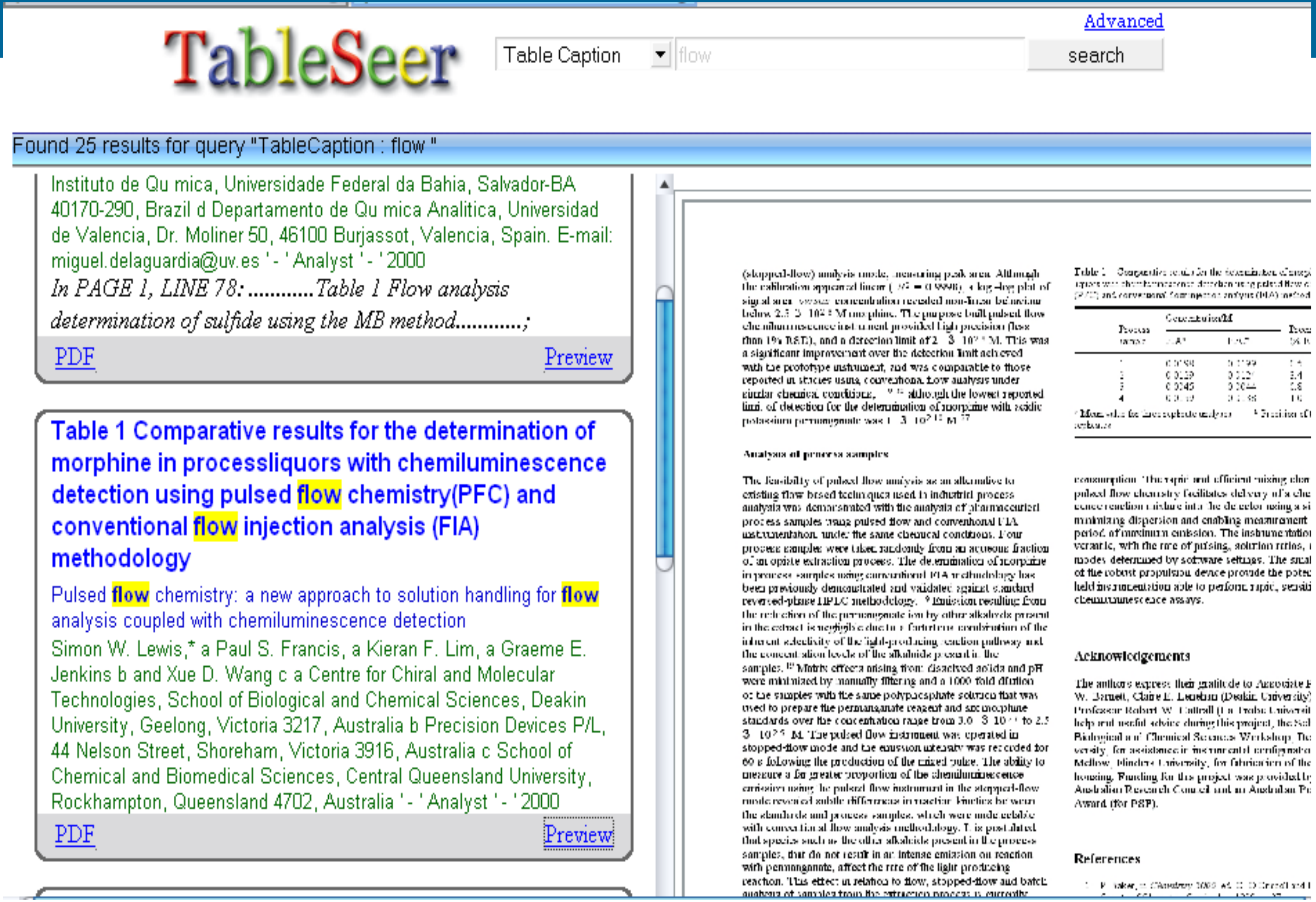
THE ANALYST
FULL PAPER
www.rsc.org/analyst

Accelerator mass spectrometry (AMS) was used to measure ^{237}Np in environmental water samples extracted from Irish Sea sediments. The samples were of limited volume (~700 ml) and of low activity (0.06–0.79 mBq l−1;

College of
Information Sciences and Technology

TableSeer

Beta online working design of a table search engine



Advanced

TableSeer

Table Caption ▾ | flow | search

Found 25 results for query "TableCaption : flow "

Instituto de Qu mica, Universidade Federal da Bahia, Salvador-BA 40170-290, Brazil d Departamento de Qu mica Analitica, Universidad de Valencia, Dr. Moliner 50, 46100 Burjassot, Valencia, Spain. E-mail: miguel.delaguardia@uv.es '- ' Analyst ' - ' 2000

In PAGE 1, LINE 78: ....Table 1 Flow analysis determination of sulfide using the MB method............;

PDF | Preview

Table 1 Comparative results for the determination of morphine in processliquors with chemiluminescence detection using pulsed flow chemistry(PFC) and conventional flow injection analysis (FIA) methodology

Pulsed flow chemistry: a new approach to solution handling for flow analysis coupled with chemiluminescence detection

Simon W. Lewis,* a Paul S. Francis, a Kieran F. Lim, a Graeme E. Jenkins b and Xue D. Wang c a Centre for Chiral and Molecular Technologies, School of Biological and Chemical Sciences, Deakin University, Geelong, Victoria 3217, Australia b Precision Devices P/L, 44 Nelson Street, Shoreham, Victoria 3916, Australia c School of Chemical and Biomedical Sciences, Central Queensland University, Rockhampton, Queensland 4702, Australia ' - ' Analyst ' - ' 2000

PDF | Preview

(stopped-flow) analysis mode, measuring peak area. Although the calibration appeared linear (r^2 = 0.9998), a log–log plot of signal area *versus* concentration revealed non-linear behaviour below 2.5 3 10?8 M morphine. The purpose built pulsed flow chemiluminescence instrument provided high precision (less than 1% R.S.D.), and a detection limit of 2 3 10?9 M. This was a significant improvement over the detection limit achieved with the prototype instrument, and was comparable to those reported in studies using conventional flow analysis under similar chemical conditions,[9,16] although the lowest reported limit of detection for the determination of morphine with acidic potassium permanganate was 1 3 10?12 M.[17]

Analysis of process samples

The feasibility of pulsed flow analysis as an alternative to existing flow based techniques used in industrial process analysis was demonstrated with the analysis of pharmaceutical process samples using pulsed flow and conventional FIA instrumentation under the same chemical conditions. Four process samples were taken randomly from an aqueous fraction of an opiate extraction process. The determination of morphine in process samples using conventional FIA methodology has been previously demonstrated and validated against standard reversed-phase HPLC methodology.[9] Emission resulting from the reduction of the permanganate ion by other alkaloids present in the extract is negligible due to a fortuitous combination of the inherent selectivity of the light-producing reaction pathway and the concentration levels of the alkaloids present in the samples.[18] Matrix effects arising from dissolved solids and pH were minimized by manually filtering and a 1000 fold dilution of the samples with the same polyphosphate solution that was used to prepare the permanganate reagent and morphine standards over the concentration range from 3.0 3 10?? to 2.5 3 10?5 M. The pulsed flow instrument was operated in stopped-flow mode and the emission intensity was recorded for 60 s following the production of the mixed pulse. The ability to measure a far greater proportion of the chemiluminescence emission using the pulsed flow instrument in the stopped-flow mode revealed subtle differences in reaction kinetics between the standards and process samples, which were undetectable with conventional flow analysis methodology. It is postulated that species such as the other alkaloids present in the process samples, that do not result in an intense emission on reaction with permanganate, affect the rate of the light producing reaction. This effect in relation to flow, stopped-flow and batch analysis of samples from the extraction process is currently

Table 1 Comparative results for the determination of morphi
liquors with chemiluminescence detection using pulsed flow c
(PFC) and conventional flow injection analysis (FIA) method

Process sample	Concentration/M PFC^a	 FIA^b	Preci (% R
1	[illegible]	[illegible]	1.4
2	0.0129	0.0127	3.4
3	0.0045	0.0044	0.8
4	[illegible]	[illegible]	1.0

^{a} Mean value for three replicate analyses. ^{b} Precision of t
replicates.

consumption. The rapid and efficient mixing chem
pulsed flow chemistry facilitates delivery of a che
escence reaction mixture into the detector using a si
minimizing dispersion and enabling measurement
period of maximum emission. The instrumentation
versatile, with the rate of pulsing, solution ratios, a
modes determined by software settings. The smal
of the robust propulsion device provide the poten
held instrumentation able to perform rapid, sensiti
chemiluminescence assays.

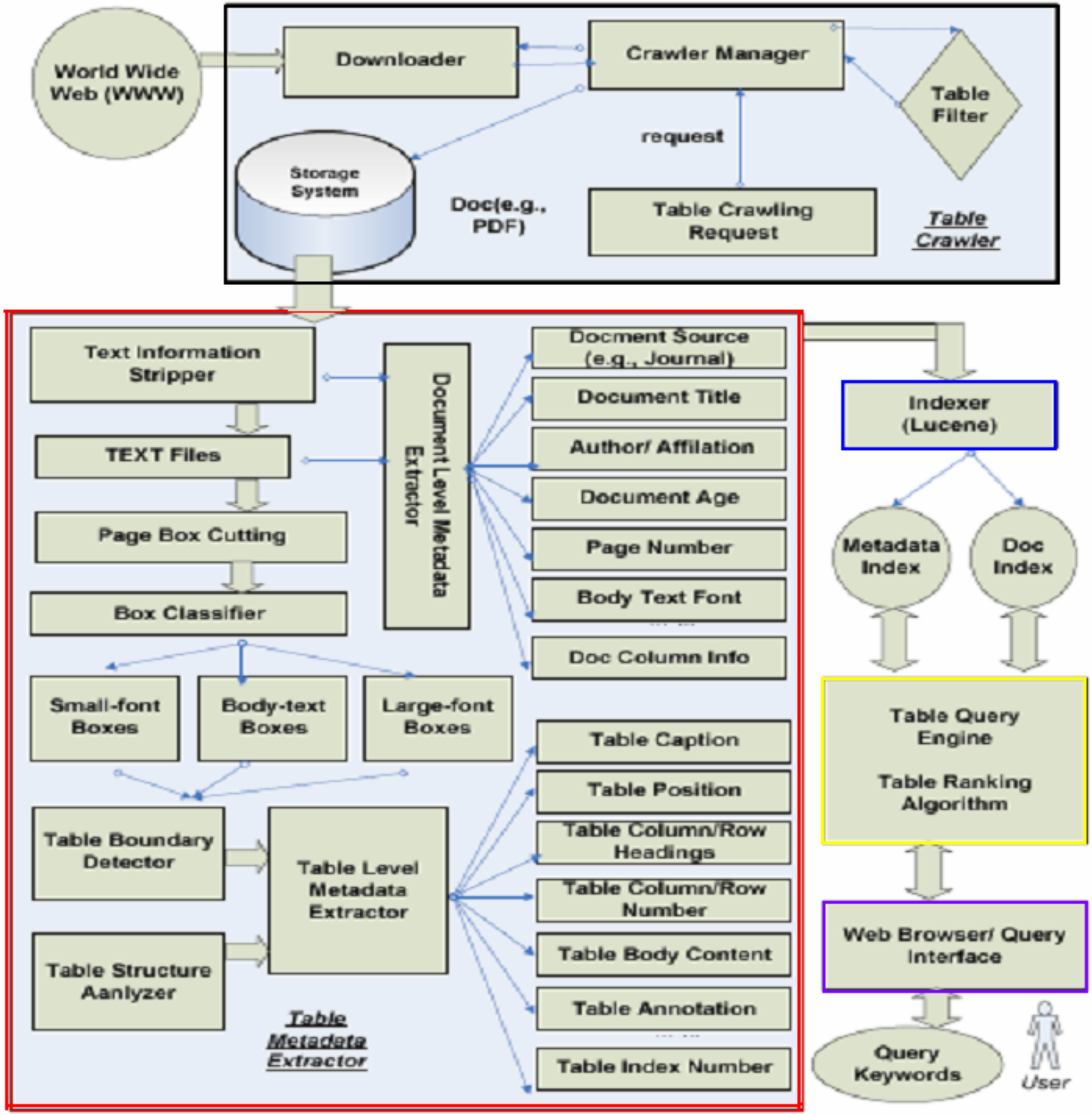
Acknowledgements

The authors express their gratitude to Associate P
W. Barnett, Claire E. Lenehan (Deakin University)
Professor Robert W. Cattrall (La Trobe Universit
help and useful advice during this project, the Sch
Biological and Chemical Sciences Workshop, De
versity, for assistance in instrumental configuratio
Mellow, Flinders University, for fabrication of the
housing. Funding for this project was provided by
Australian Research Council and an Australian Po
Award (for PSF).

References

1 P. Baker, [illegible]

TableSeer System Architecture



Sample Table Metadata Extracted File

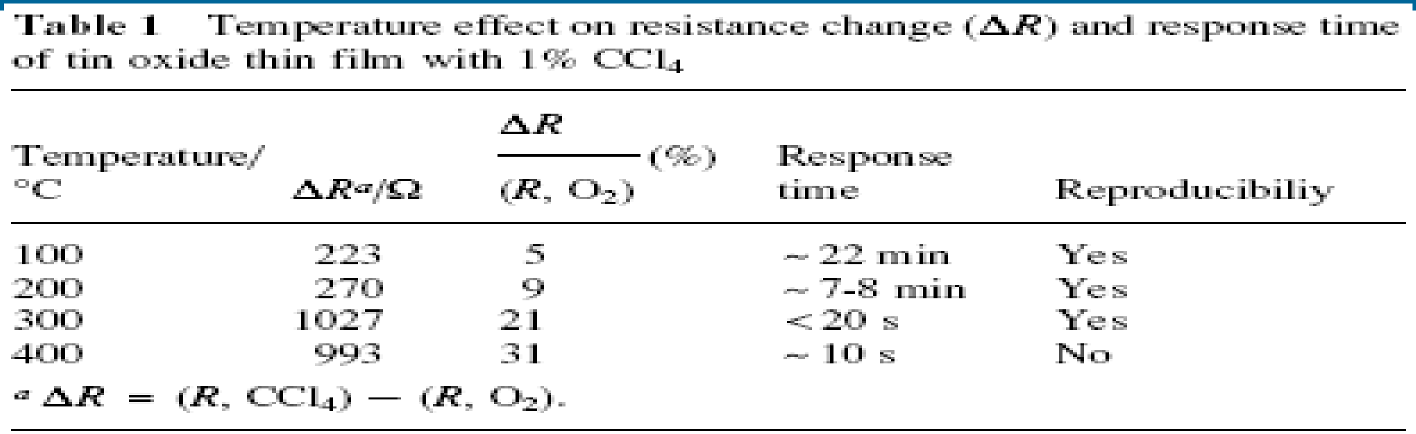


Table 1 Temperature effect on resistance change (ΔR) and response time of tin oxide thin film with 1% CCl_4

Temperature/ °C	$\Delta R^a/\Omega$	$\frac{\Delta R}{(R, O_2)}$ (%)	Response time	Reproducibiliy
100	223	5	~ 22 min	Yes
200	270	9	~ 7-8 min	Yes
300	1027	21	< 20 s	Yes
400	993	31	~ 10 s	No

[a] $\Delta R = (R, CCl_4) - (R, O_2)$.

- <Table>
- <DocumentOrigin>Analyst</DocumentOrigin>
- <DocumentName>b006011i.pdf</DocumentName>
- <Year>2001</Year>
- <DocumentTitle>Detection of chlorinated methanes by tin oxide gas sensors </DocumentTitle>
- <Author>Sang Hyun Park, a ? Young-Chan Son, a Brenda R . Shaw, a Kenneth E. Creasy,* b and Steven L. Suib* acd a Department of Chemistry, U-60, University of Connecticut, Storrs, C T 06269-3060</Author>
- <TheNumOfCiters></TheNumOfCiters>
- <Citers></Citers>
- <TableCaption>Table 1 Temperature effect o n r esistance change (D R) and response timeof tin oxide thin film with 1 % C Cl 4</TableCaption>
- <TableColumnHeading>D R Temperature/ ¡ã C D R a / W (R ,O 2) (%) R esponse time Reproducibiliy </TableColumnHeading>
- <TableContent>100 223 5 ~ 22 min Yes 200 270 9 ~ 7-8 min Yes 300 1027 21 < 2 0 s Yes 400 993 31 ~ 1 0 s No </TableContent>
- <TableFootnote> a D R =(R , CCl 4) - (R ,O 2). </TableFootnote>
- <ColumnNum>5</ColumnNum>
- <TableReferenceText>In page 3, line 11, … Film responses to 1% CCl4 at different temperatures are summarized in Table 1……</TableReferenceText>
- <PageNumOfTable>3</PageNumOfTable>
- <Snapshot>b006011i/b006011i_t1.jpg</Snapshot>
- </Table>

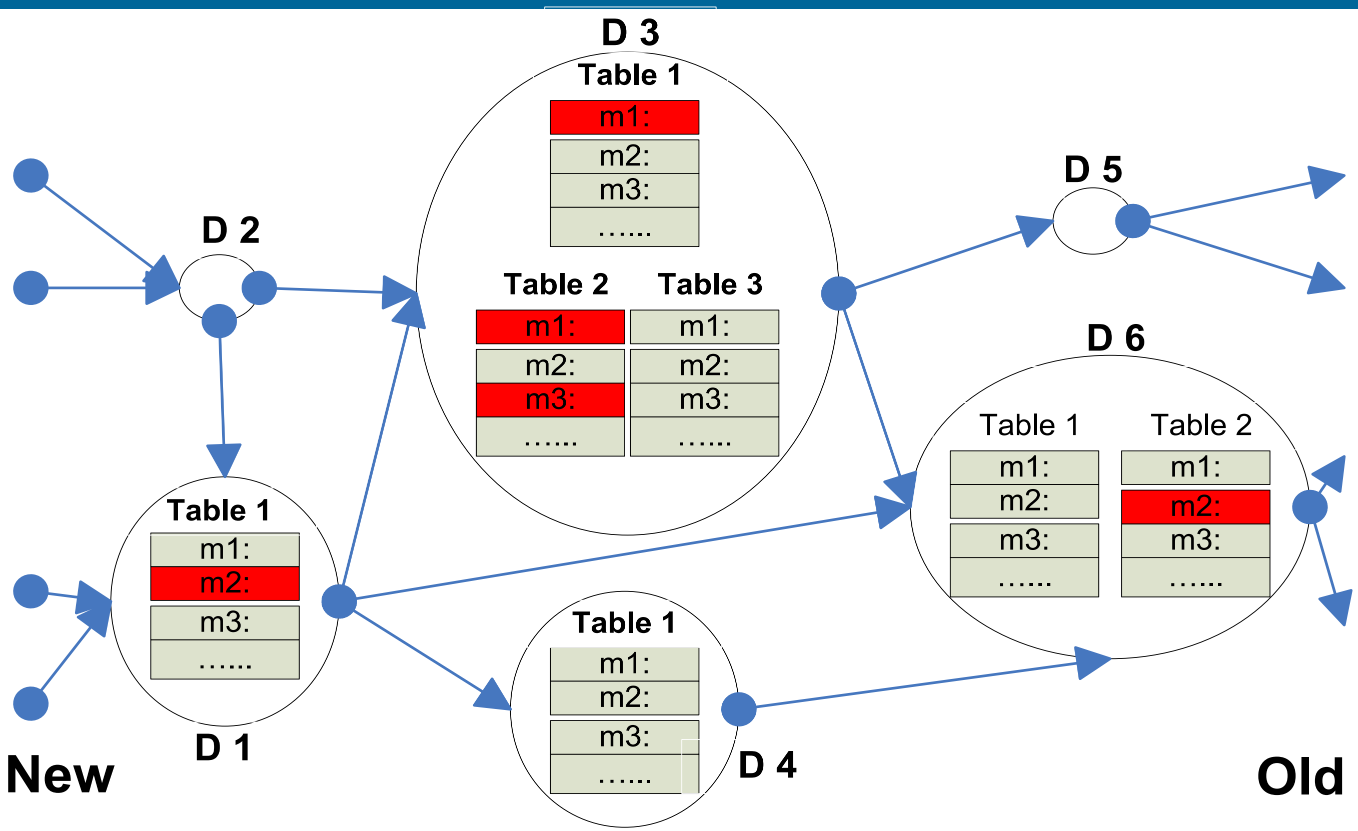
TableRank

- Rank tables by rating the <query, table> pairs, instead of the <query, document> pairs: preventing a lot of false positive hits for table search, which frequently occur in current web search engines
- The similarity between a <table, query> pair: the cosine of the angle between vectors

$$sim(tb_j, Q) = cos(tb_j, Q) = \frac{\sum_{i=1}^{s} w_{i,j,k} w_{i,q,k}}{|tb_j||Q|}$$

- Tailored term vector space => table vectors:
  - Query vectors and table vectors, instead of document vectors

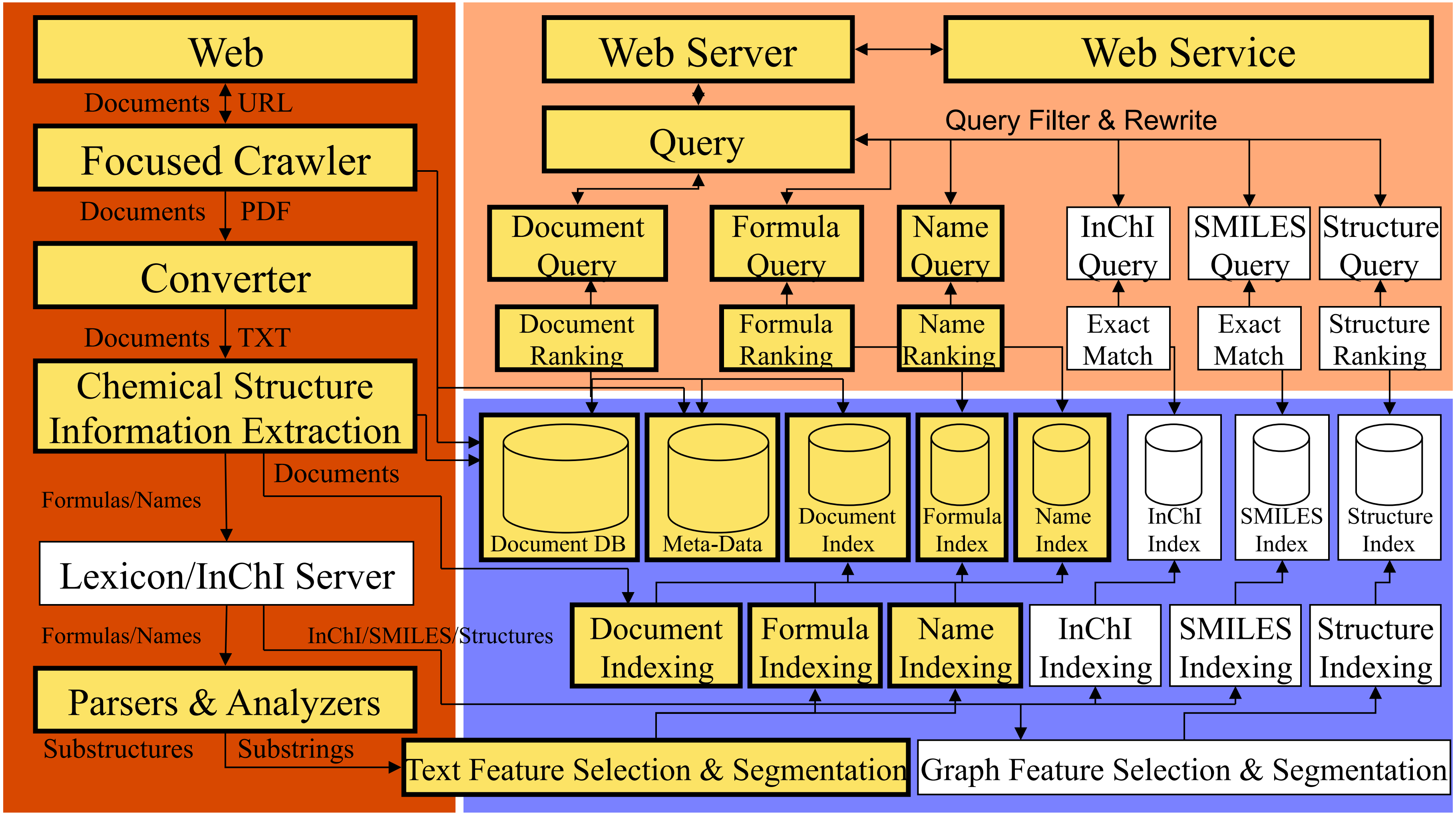
An Example of the Citation Network



Challenges

- Identification of cells
  - Cells getting fused
- Irregularly shaped tables
- Horizontal letters
- Identify columns
  - Sometimes columns and rows are fused
  - Define heuristics to detect such cases
- Identifying what the columns signify
- units of the columns

Proposed Architecture and Framework



Digital Libraries

- Connecting data and digital documents
  - Mining legacy data from tables in digital documents
  - Data submission system where creator can link data and documents
- Use ORE model

Data-Document IntegrationNeeds

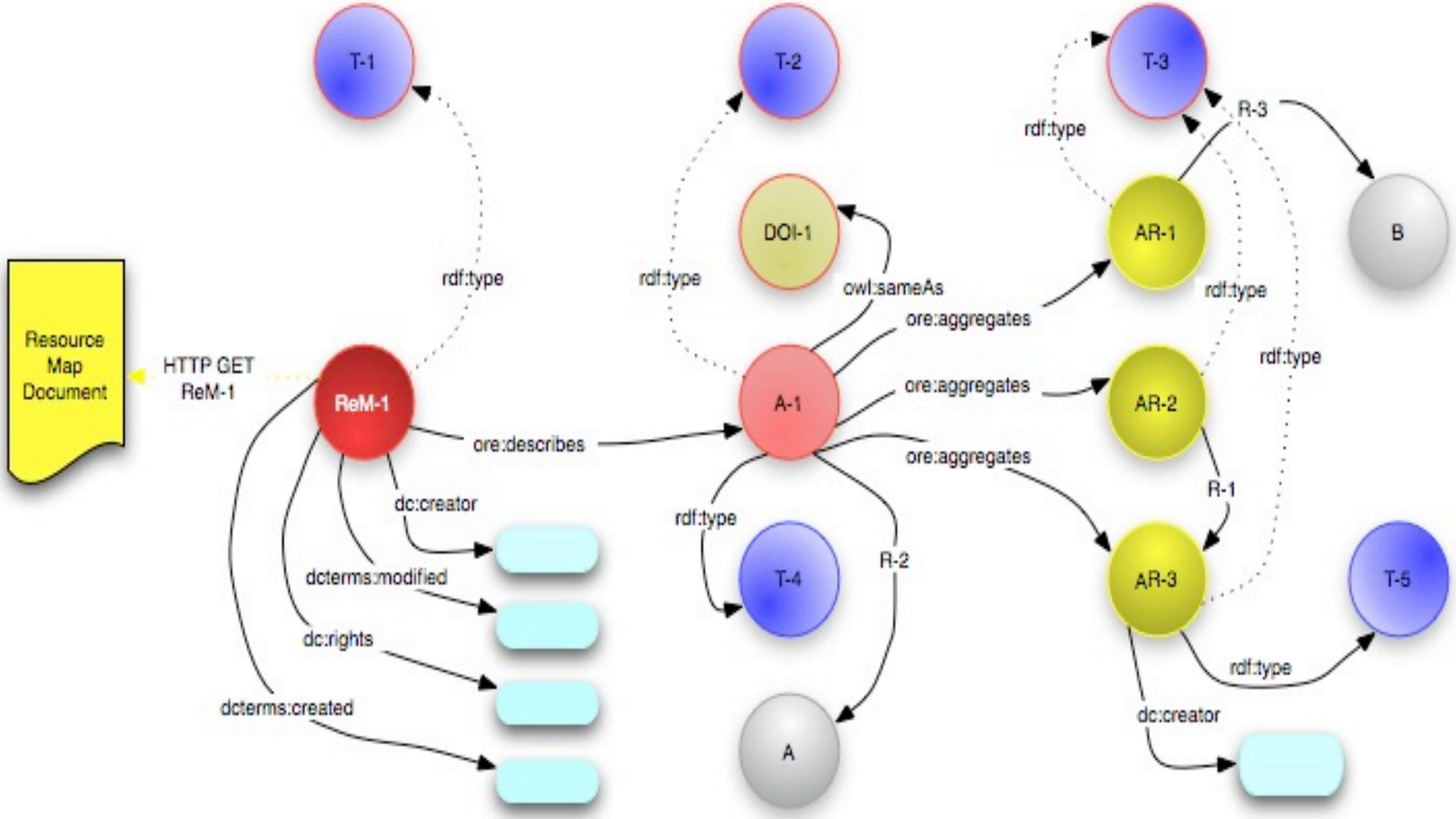
- Queries:
  - Get me the data on the dissolution rate of kaolinite
  - Get me the data & experimental results dissolution rate of kaolinite reported in papers referred to by paper x.

- Link information from tables in databases to documents
- Table-level Links
  - All data from a table from a single article
- Row-level Links
  - Data in table linking to multiple articles

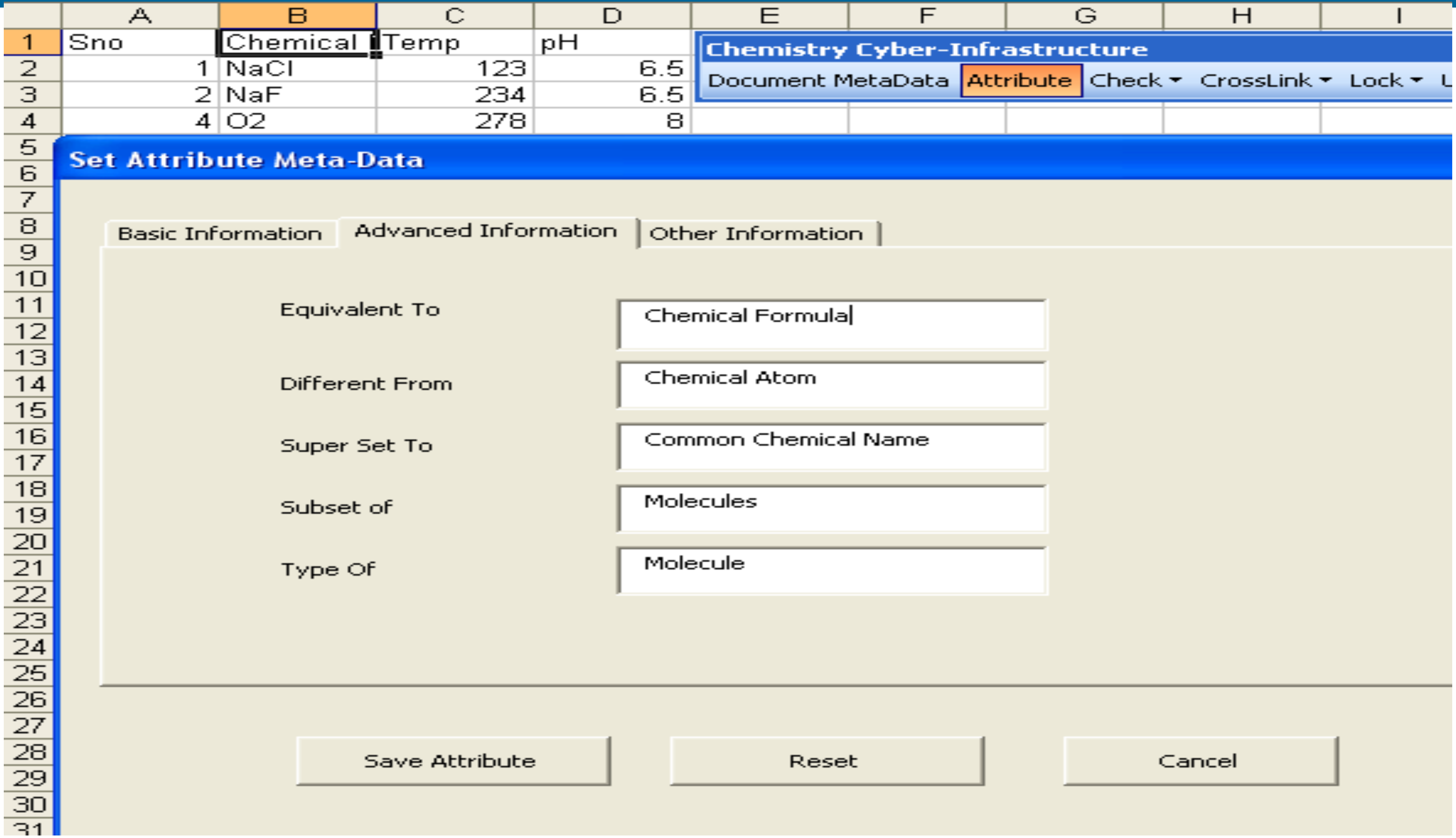
Use ORE

- Dataset – resource
- Aggregates of linked resources
  - Database
    - Published
    - Raw
  - Article
    - Preprints
  - Supplementary notes
- Resource Map

Digital Objects



Linking to Ontology



	A	B	C	D
1	Sno	Chemical	Temp	pH
2	1	NaCl	123	6.5
3	2	NaF	234	6.5
4	4	O2	278	8

Set Attribute Meta-Data

Equivalent To: Chemical Formula

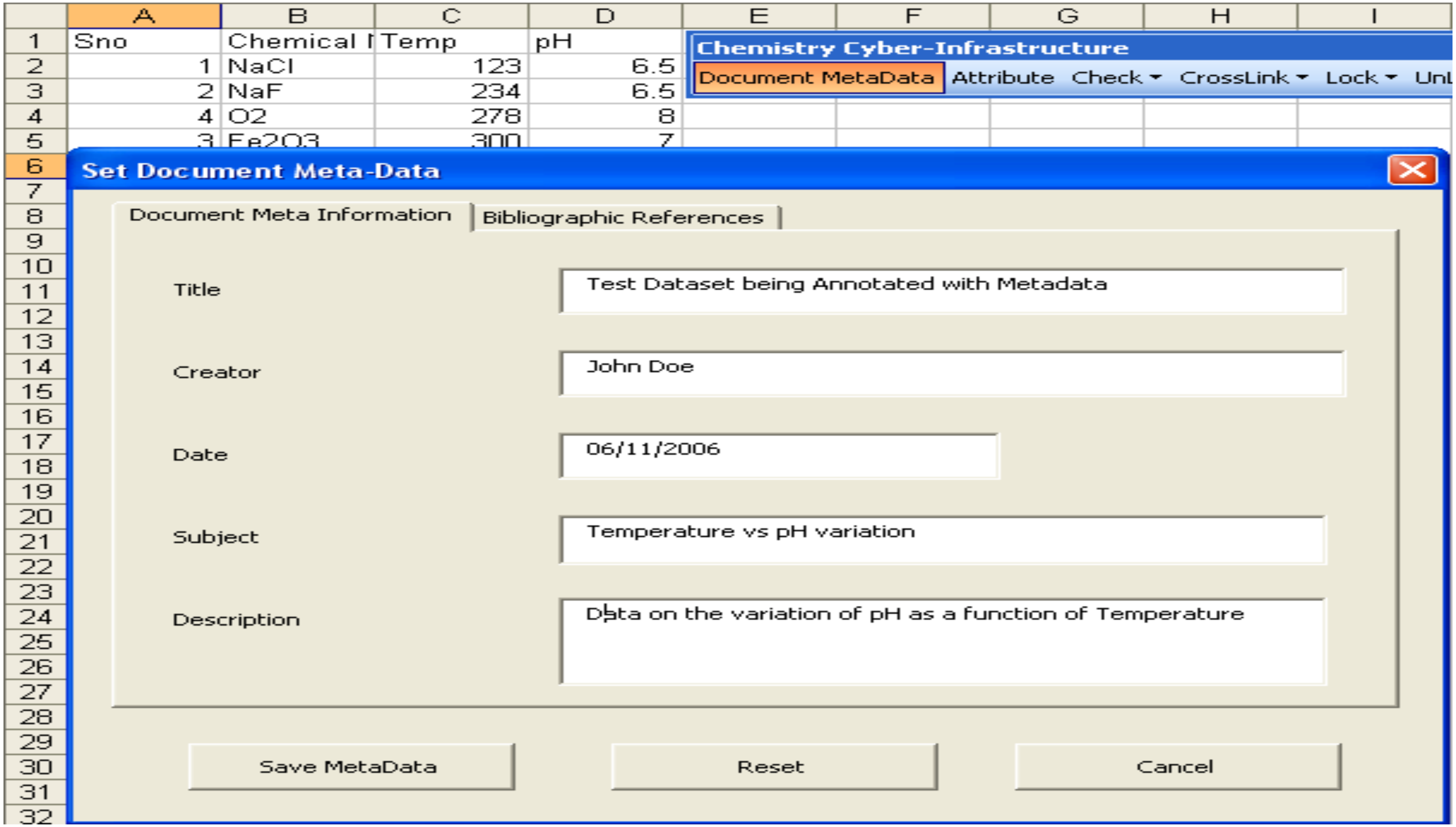
Different From: Chemical Atom

Super Set To: Common Chemical Name

Subset of: Molecules

Type Of: Molecule

Data Annotation



	A	B	C	D
1	Sno	Chemical	Temp	pH
2	1	NaCl	123	6.5
3	2	NaF	234	6.5
4	4	O2	278	8
5	3	Fe2O3	300	7

Title: Test Dataset being Annotated with Metadata

Creator: John Doe

Date: 06/11/2006

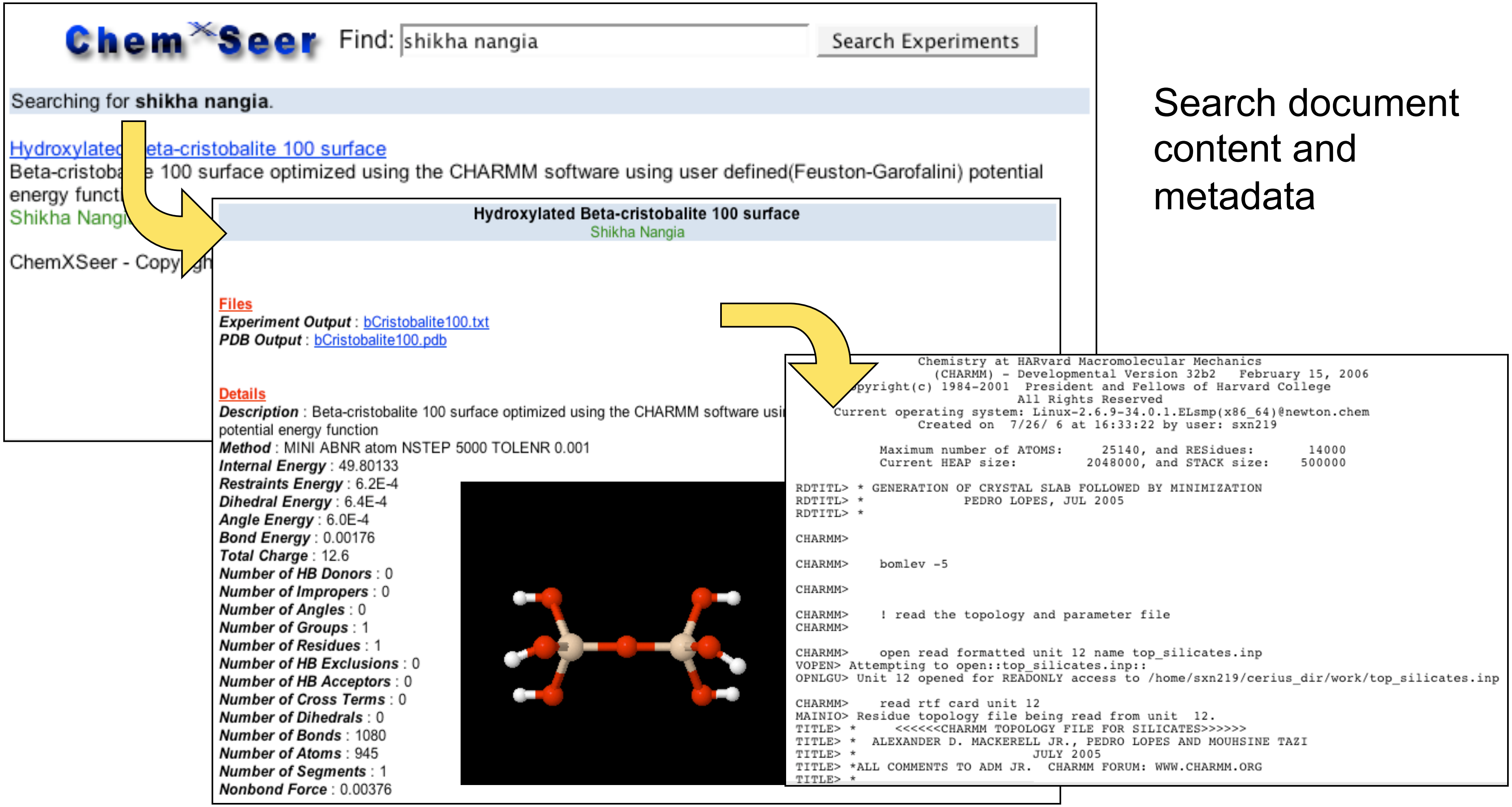
Subject: Temperature vs pH variation

Description: Data on the variation of pH as a function of Temperature

Ontology-based Mediated Search

- Ontologies used to express vocabularies
  - Mediate federated databases
  - Mediated search
    - Multiple data formats
      - Gaussian, CHARMM, Excel, XML, ASCII …

Data Search Example in Chem$_X$Seer



Searching for **shikha nangia**.

Hydroxylated eta-cristobalite 100 surface
Beta-cristoba e 100 surface optimized using the CHARMM software using user defined(Feuston-Garofalini) potential energy funct
Shikha Nang

ChemXSeer - Copy gh

Hydroxylated Beta-cristobalite 100 surface
Shikha Nangia

Files
Experiment Output : bCristobalite100.txt
PDB Output : bCristobalite100.pdb

Details
Description : Beta-cristobalite 100 surface optimized using the CHARMM software usi
potential energy function
Method : MINI ABNR atom NSTEP 5000 TOLENR 0.001
Internal Energy : 49.80133
Restraints Energy : 6.2E-4
Dihedral Energy : 6.4E-4
Angle Energy : 6.0E-4
Bond Energy : 0.00176
Total Charge : 12.6
Number of HB Donors : 0
Number of Impropers : 0
Number of Angles : 0
Number of Groups : 1
Number of Residues : 1
Number of HB Exclusions : 0
Number of HB Acceptors : 0
Number of Cross Terms : 0
Number of Dihedrals : 0
Number of Bonds : 1080
Number of Atoms : 945
Number of Segments : 1
Nonbond Force : 0.00376

```
                 Chemistry at HARvard Macromolecular Mechanics
                   (CHARMM) - Developmental Version 32b2   February 15, 2006
        opyright(c) 1984-2001  President and Fellows of Harvard College
                              All Rights Reserved
      Current operating system: Linux-2.6.9-34.0.1.ELsmp(x86_64)@newton.chem
                 Created on  7/26/ 6 at 16:33:22 by user: sxn219

            Maximum number of ATOMS:     25140, and RESidues:       14000
            Current HEAP size:         2048000, and STACK size:    500000

RDTITL> * GENERATION OF CRYSTAL SLAB FOLLOWED BY MINIMIZATION
RDTITL> *            PEDRO LOPES, JUL 2005
RDTITL> *

CHARMM>

CHARMM>    bomlev -5

CHARMM>

CHARMM>    ! read the topology and parameter file
CHARMM>

CHARMM>    open read formatted unit 12 name top_silicates.inp
VOPEN> Attempting to open::top_silicates.inp::
OPNLGU> Unit 12 opened for READONLY access to /home/sxn219/cerius_dir/work/top_silicates.inp

CHARMM>    read rtf card unit 12
MAINIO> Residue topology file being read from unit  12.
TITLE> *     <<<<<<CHARMM TOPOLOGY FILE FOR SILICATES>>>>>>
TITLE> *  ALEXANDER D. MACKERELL JR., PEDRO LOPES AND MOUHSINE TAZI
TITLE> *                        JULY 2005
TITLE> *ALL COMMENTS TO ADM JR.  CHARMM FORUM: WWW.CHARMM.ORG
TITLE> *
```

Search document content and metadata

Utilizing the Social Network

- Automatic discovery of people who have similar interests (optional feature)
  - Who reads the same people/papers?
  - Who cites the same people/papers?
  - Who produces similar work?
  - Who has similar browsing habits?
  - Zhou, Manavoglu, et al, 2006; Zhou, Ji, et al, 2006
- Data traditionally used for collaborative filtering
  - Instead: connect people, build community awareness
  - Discover new people/trends of interest
- Shy users may cloak their profiles, if desired

Personalization

- MyCiteSeerX

Conclusion

- Repository for data and documents
  - Linking data to documents
- Intelligent tools for data extraction
  - Data analysis capabilities can be built in
- Advanced search capabilities
  - Chemically aware search

Thanks!

- http://chemxseer.ist.psu.edu
- http://citeseerx.ist.psu.edu

- pmitra@ist.psu.edu
- giles@ist.psu.edu

Conclusion

- Repository for data and documents
  - Linking data to documents
- Intelligent tools for data extraction
  - Data analysis capabilities can be built in
- Advanced search capabilities
  - Chemically aware search

Appeal

- Please submit your data and documents!!!
- The portal will not succeed without data

Needs

- What do you need?
- pmitra@ist.psu.edu