19th International Computer Science and Engineering Conference (ICSEC)
Chiang Mai, Thailand, 23-26 November, 2015

NMRexSeer: Metadata Extraction and Search for Large Scale Nuclear Magnetic Resonance (NMR) Experimental Data

Suppawong Tuarob
Information and Communication Technology
Mahidol University, Thailand
suppawong.tua@mahidol.ac.th

Kevin A Glass
The William R. Wiley Environmental
Molecular Sciences Laboratory
Pacific Northwest National Laboratory, USA
kevin.glass@pnnl.gov

C. Lee Giles
Information Sciences and Technology
Pennsylvania State University, USA
giles@ist.psu.edu

***Abstract*—Sciences have become both complex and demanding for cutting-edge technology and resources to perform experiments. Since 1997, the Environmental Molecular Sciences Laboratory (EMSL) has served as a user facility housing resources for global scientists to perform experiments necessary to their research. Overtime, the generated data has become both massive and redundant. To encourage better management and reuse of such experimental data, MyEMSL has emerged as an in-house centralized data management tool that collects and distributes data from the experiments at EMSL. Nuclear Magnetic Resonance Spectroscopy (NMR) is one of the major experiment resources that EMSL houses. We discuss *NMRexSeer*, a proposed digital library system that automatically extracts and indexes NMR specific metadata from NMR experimental data packages. The system also generates visualized previews and provides a search interface for easy access and discovery of desired data.**

I. Introduction

The William R. Wiley Environmental Molecular Sciences Laboratory (EMSL) [1] is a national user facility which provides experimental and computational resources to address the environmental molecular science challenges facing the Department of Energy (DOE) and the nation. EMSL is funded by the DOE's Office of Biological and Environmental Research to support its mission to provide innovative solutions to the nation's environmental and energy production challenges in areas such as atmospheric aerosols, feedstocks, global carbon cycling, biogeochemistry, subsurface science, and energy materials.

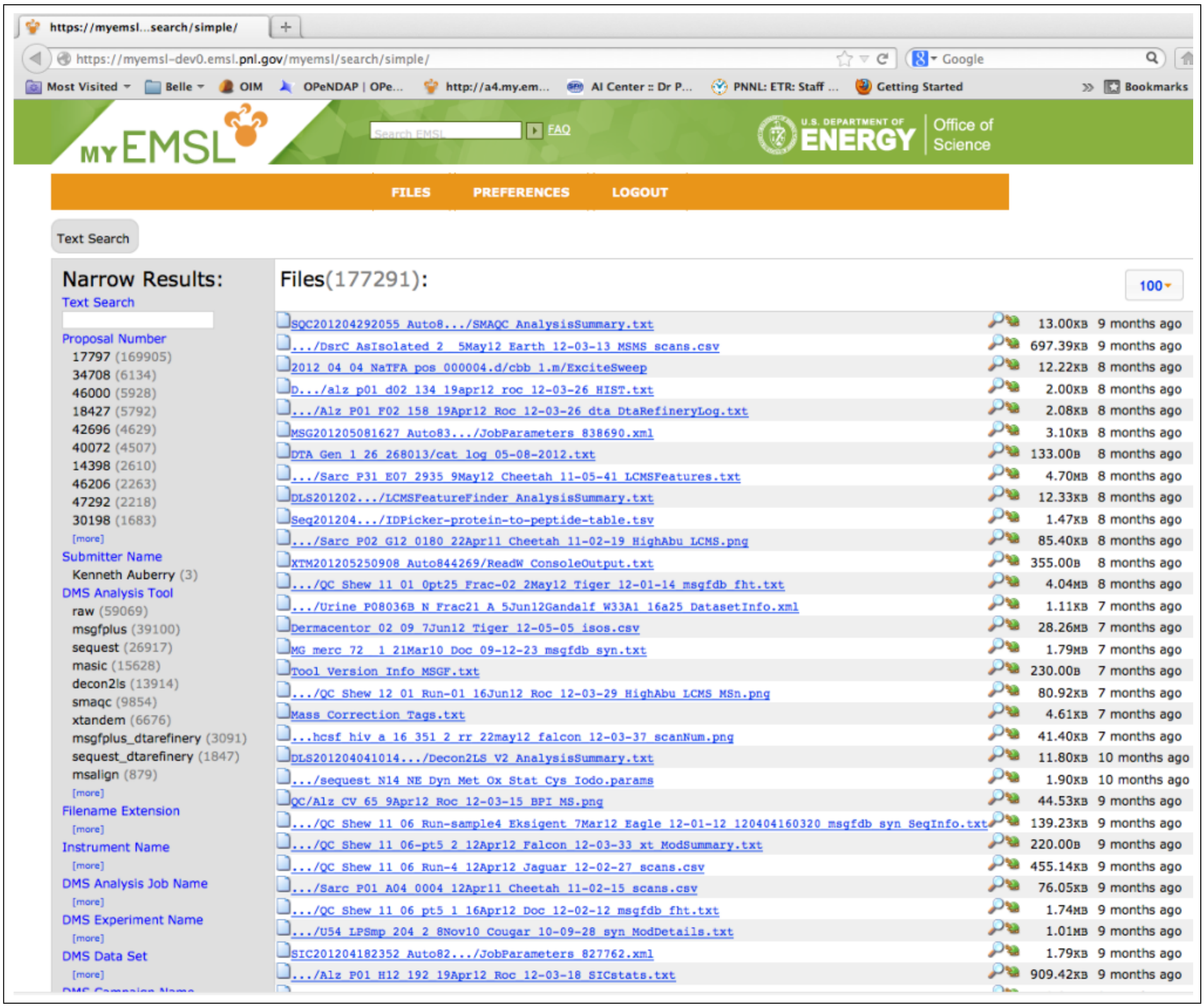
EMSL offers the global scientific community a range of capabilities and expertise. Access to EMSL's capabilities is gained through a peer-reviewed proposal process. If a proposal is accepted and the scientist publishes in the open literature, there typically is no charge for using the EMSL instrumentation and capabilities. With such growing interest from the scientific community, EMSL is currently facing the following problems:

1) **Resource Over-usage**. EMSL hosts cutting-edge biology and chemistry experimental resources that attract an enormous amount of scientists needing to access such resources. Oftentimes, the resources are limited and only allocated to those whose proposals are approved.
2) **Remote Data Access and Sharing**. Scientists who do not have physical access to the facility may inquire that the experiments be done remotely on their behalf. However, sending the experimental data back to the requesting scientists can be problematic since there is not one conventional method for delivering digitalized data. Moreover, most data packages are bigger than allowed email attachment sizes, making transporting data packages via emails not feasible.
3) **Massive Data**. After each experiment, EMSL stores digitalized data as backups and for knowledge mining applications. In 2013, Cowley reported 6TB of digitalized data was produced daily [2]. He also projected that the data produced daily at EMSL will become as large as 100TB. This calls for a system that not only efficiently manages this ever increasing amount of data, but also makes it readily accessible.
4) **Redundant Data**. Oftentimes, scientists perform the same sets of experiments that have been done before, resulting in not only inefficient use of resources, but also redundant data in the storage.

To mitigate such problems, EMSL has recently initiated *MyEMSL*, a data management system that serves as a portal for accessing all the experimental data produced at EMSL. MyEMSL collects experimental data either manually or automatically. In either case, data collection includes the collection of limited metadata. The data from an experiment is referred to as a *package* and is stored in a zip file (most instruments generate multiple files for each experiment). MyEMSL then stores all the packages into an archive and makes them searchable and accessible via a search interface as depicted in Figure 1. Even though MyEMSL also collects the metadata from the packages to enable searching, the collected metadata is not experiment specific, only containing general experiment information such as package names, submitters' names, submission dates, instruments used to perform the experiments, etc. Hence, a better metadata extraction that provides specific information about the experiment technique could prove to facilitate search and discovery.

Over the past fifty years, nuclear magnetic resonance spectroscopy, commonly referred to as *NMR*, has become the preeminent technique for determining the structure of organic

978-1-4673-7825-3/15/$31.00 © 2015 IEEE



Narrow Results:

Text Search

Proposal Number
17797 (169905)
34708 (6134)
46000 (5928)
18427 (5792)
42696 (4629)
40072 (4507)
14398 (2610)
46206 (2263)
47292 (2218)
30198 (1683)
[more]

Submitter Name
Kenneth Auberry (3)

DMS Analysis Tool
raw (59069)
msgfplus (39100)
sequest (26917)
masic (15628)
decon2ls (13914)
smaqc (9854)
xtandem (6676)
msgfplus_dtarefinery (3091)
sequest_dtarefinery (1847)
msalign (879)
[more]

Filename Extension
[more]

Instrument Name
[more]

DMS Analysis Job Name
[more]

DMS Experiment Name
[more]

DMS Data Set
[more]

Files(177291):

SQC201204292055 Auto8.../SMAQC AnalysisSummary.txt	13.00KB	9 months ago
.../DsrC AsIsolated 2 5May12 Earth 12-03-13 MSMS scans.csv	697.39KB	9 months ago
2012 04 04 NaTFA pos 000004.d/cbb 1.m/ExciteSweep	12.22KB	8 months ago
D.../alz p01 d02 134 19apr12 roc 12-03-26 HIST.txt	2.00KB	8 months ago
.../Alz P01 F02 158 19Apr12 Roc 12-03-26 dta DtaRefineryLog.txt	2.08KB	8 months ago
MSG201205081627 Auto83.../JobParameters 838690.xml	3.10KB	8 months ago
DTA Gen 1 26 268013/cat log 05-08-2012.txt	133.00B	8 months ago
.../Sarc P31 E07 2935 9May12 Cheetah 11-05-41 LCMSFeatures.txt	4.70MB	8 months ago
DLS201202.../LCMSFeatureFinder AnalysisSummary.txt	12.33KB	8 months ago
Seq201204.../IDPicker-protein-to-peptide-table.tsv	1.47KB	8 months ago
.../Sarc P02 G12 0180 22Apr11 Cheetah 11-02-19 HighAbu LCMS.png	85.40KB	8 months ago
XTM201205250908 Auto844269/ReadW ConsoleOutput.txt	355.00B	8 months ago
.../QC Shew 11 01 0pt25 Frac-02 2May12 Tiger 12-01-14 msgfdb fht.txt	4.04MB	8 months ago
.../Urine P08036B N Frac21 A 5Jun12Gandalf W33A1 16a25 DatasetInfo.xml	1.11KB	7 months ago
Dermacentor 02 09 7Jun12 Tiger 12-05-05 isos.csv	28.26MB	7 months ago
MG merc 72 1 21Mar10 Doc 09-12-23 msgfdb syn.txt	1.79MB	7 months ago
Tool Version Info MSGF.txt	230.00B	7 months ago
.../QC Shew 12 01 Run-01 16Jun12 Roc 12-03-29 HighAbu LCMS MSn.png	80.92KB	7 months ago
Mass Correction Tags.txt	4.61KB	7 months ago
...hcsf hiv a 16 351 2 rr 22may12 falcon 12-03-37 scanNum.png	41.40KB	7 months ago
DLS201204041014.../Decon2LS V2 AnalysisSummary.txt	11.80KB	10 months ago
.../sequest N14 NE Dyn Met Ox Stat Cys Iodo.params	1.90KB	10 months ago
QC/Alz CV 65 9Apr12 Roc 12-03-15 BPI MS.png	44.53KB	9 months ago
.../QC Shew 11 06 Run-sample4 Eksigent 7Mar12 Eagle 12-01-12 120404160320 msgfdb syn SeqInfo.txt	139.23KB	9 months ago
.../QC Shew 11 06-pt5 2 12Apr12 Falcon 12-03-33 xt ModSummary.txt	220.00B	9 months ago
.../QC Shew 11 06 Run-4 12Apr12 Jaguar 12-02-27 scans.csv	455.14KB	9 months ago
.../Sarc P01 A04 0004 12Apr11 Cheetah 11-02-15 scans.csv	76.05KB	9 months ago
.../QC Shew 11 06 pt5 1 16Apr12 Doc 12-02-12 msgfdb fht.txt	1.74MB	9 months ago
.../U54 LPSmp 204 2 8Nov10 Cougar 10-09-28 syn ModDetails.txt	1.01MB	9 months ago
SIC201204182352 Auto82.../JobParameters 827762.xml	1.79KB	9 months ago
.../Alz P01 H12 192 19Apr12 Roc 12-03-18 SICstats.txt	909.42KB	9 months ago

Fig. 1. Sample screen shot of MyEMSL data management system, taken from [2].

compounds [3]. EMSL houses over 11 NMR spectrometers from two major vendors (Bruker [4] and Varian Inc. [5]) [6]. Each NMR experiment involves reading both time-domain and frequency-domain spectra of a test substance. The same setting is usually run multiple times to filter out noise in the experiment results.

In this paper, *NMRexSeer* is described. The system extends MyEMSL by extracting NMR specific metadata from NMR experimental data packages. The system then indexes the extracted metadata using Sphinx Search Server [7], which makes the metadata searchable via the NMRexSeer search interface. The proposed system could prove to be useful to scientists searching for raw data of existing NMR experiments.

II. Related Works

Systems that extract NMR specific metadata from experimental data have not existed in the literature. Hence, only works closed to ours are discussed.

A. NMR Related Search Engines and Databases

Most existing NMR search engines and databases such as *DRESS* [8], *NMRShiftDB* [9], *MPNMR* [10], *MMCD* [11] catalog spectra of well-known substances for reference. These systems differ from the proposed NMRexSeer in two aspects:

1) These existing systems only catalog well-known, simple substances for reference, while NMRexSeer hosts NMR experiments for newly discovered/uncommon substances. NMRexSeer is hence suitable for scientists looking for experimental data of their interest substances.
2) These existing systems only present *analysed* results such as visualized spectrum images and molecular structures and statistics. However, scientists sometimes need raw experimental data to explore on their own. NMRexSeer provides both partially analysed information and raw experimental data.

B. Data Management Systems for Experimental Data

Sciences in various disciplines have become both complex and data-intensive, needing access to heterogenous data collected from multiple places, times, and thematic scales. While the needs to access such heterogenous data are apparent, the rapid expansion of experimental data, in both quantity and heterogeneity, poses huge challenges for data seekers to obtain the right information for their research. Such problems behoove tools that automatically manage, discover, and link big data from diverse sources, and present the data in the forms that are easily accessible and comprehensible.

ChemXSeer hosts over 150,000 articles from the Royal Society of Chemistry repositories [12]. ChemXSeer also extracts metadata, semantically models, and indexes these articles [13], [14]. Experimental data is published at the website and an integrated search capability allows users to search the data using its several features. Scientists can easily publish their data by uploading it onto the repository and providing optional metadata to improve the search capabilities. The data is linked to the published articles such that the user can access an article and then continue to examine the raw data after reading the article, or access a data set and then examine the article to find out the detailed experimental conditions, conclusions, discussions, and hypotheses that were validated using the experiments and their datasets. ChemXSeer was developed by the same group of computer scientists [15], [16] involving in multiple specialized search engines for document elements such as tables [17], figures [18], and algorithms [19], [20], [21], [22], [23], [24].

Recently, *DataONE*, a federated data network built to facilitate access and preservation about environmental and ecological science data across the world, has come online and is gaining increasingly popularity [25], [26], [27]. DataONE harvests metadata from different environmental data providers and makes it searchable via the search interface ONEMercury [28], built on Mercury [29], a distributed metadata management system. ONEMercury offers a full text search on the metadata records. The user can also specify the boundary of locations in which the desired data is collected or published using the interactive graphic map. At the result page, the user can choose to further filter out the results by `Member Node`, `Author`, `Project`, and `Keywords`.

With similar motivation as DataONE, *MyEMSL* has emerged as a centralized data management system that collects experimental data from various kinds of chemical and biological experiments conducted at EMSL facilities [2]. MyEMSL brings experimental capabilities (instruments, computers, etc) into a framework that provides workflow, data capture, metadata capture and generation, a central data repository, and tools for data merging, discovery, and analysis. The proposed NMRexSeer extends the capabilities of MyEMSL by specifically handling NMR-related experimental data.

III. Proposed System

The proposed system is described in Figure 2. First, MyEMSL collects experimental data packages from the experiment operators. NMRexSeer then chooses only NMR data packages to extract NMR specific metadata. The extracted metadata then is fed to the indexer, built on Sphinx Search Server, which then makes the metadata searchable via the search interface.

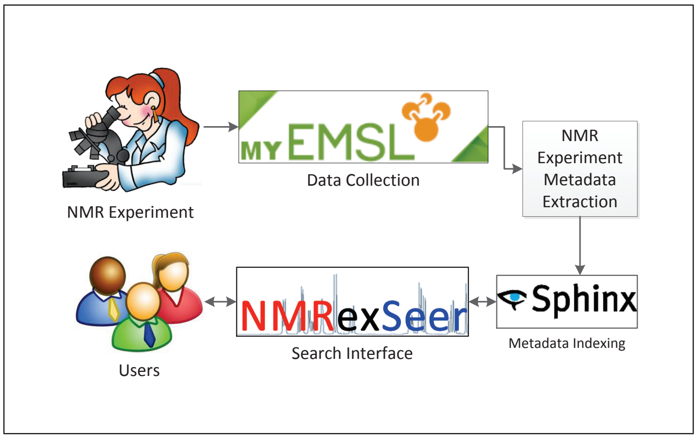


Fig. 2. Overview of *NMRexSeer* metadata extraction and search system.

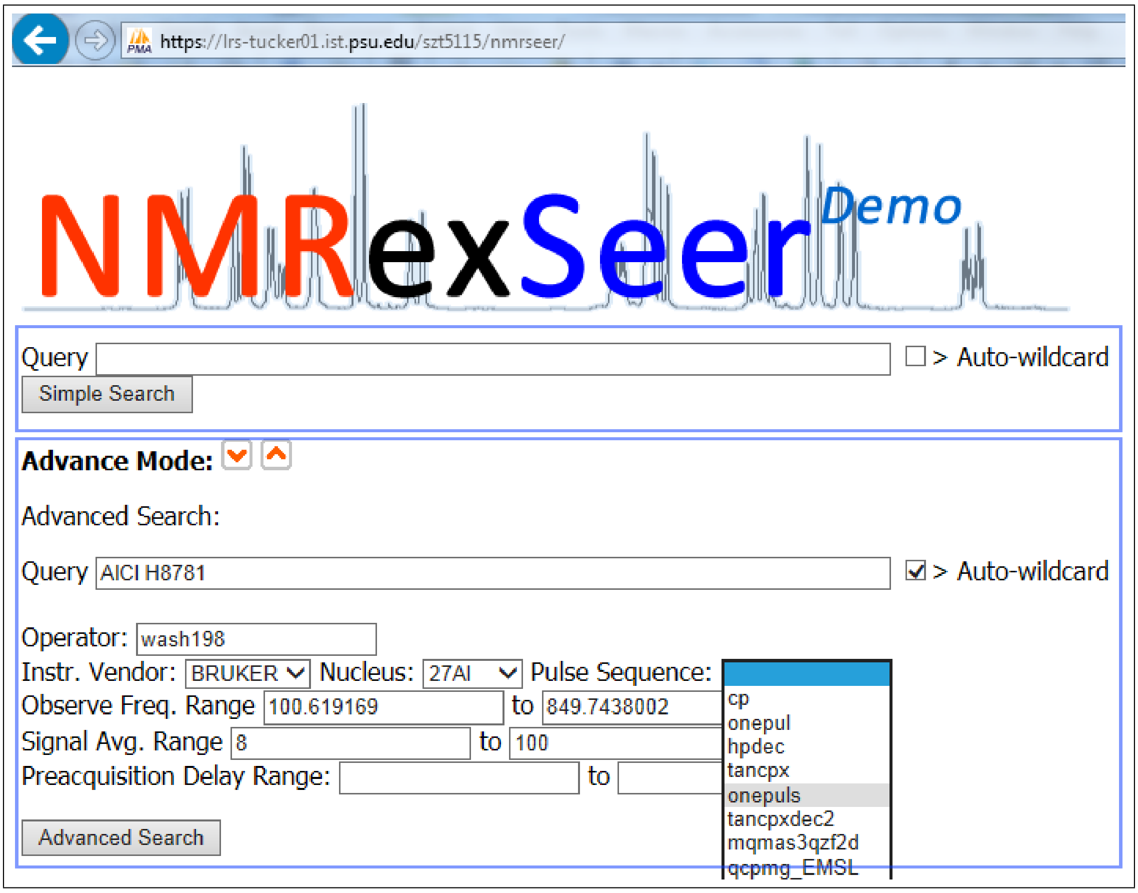
TABLE I. List of the NMR specific metadata attributes that NMRexSeer extracts from each NMR package.

Metadata	Description
Title	Experiment name
Operator	User ID of the experimenter
Device Model	Model of the NMR spectrometer
Device Name	Assigned unique name of the NMR spectrometer
Time	Date and time of experiment
Vendor	Vender of the NMR spectrometer (Bruker or Varian)
NUC1	1st nucleus
NUC2	2nd nucleus (if any)
NUC3	3rd nucleus (if any)
NUC4	4th nucleus (if any)
Pulse Seq	Pulse sequence used
Signal Avg	Signal Averaged Spectra
Preacq Delay	Delay before the first acquisition of the spectrum
Observe Freq	Observed emitted frequency of the spectra
Instrument	Instrument type (if any)
Prob Head	Prob head used
Description	Textual description provided by the experimenter
FID-Real	Real part of the free induction decay
FID-Imag	Imaginary part of the free induction decay
Spectrum	Frequency domain spectrum

A. NMR Specific Metadata Extraction

Since NMR experiment settings differ by models and vendors of the NMR spectrometers, we were advised by chemists at EMSL to collect only common metadata attributes that characterize the experiment settings and useful for searching. Table I lists all the collected metadata attributes.

Since a NMR data package available via MyEMSL is a ZIP compressed file, the NMR metadata extractor first unzips the content in the ZIP file, which mostly are a collection of ASCII files containing information about the experiment. Next, the extractor extracts the interest metadata attributes from the content. Since the experiment files are formatted differently depending on the vendors and the models of the spectrometers, a different set of regular expression rules is used for each data format. The regular expression rules are implemented according to the format specifications in the instrument user manuals. The extracted metadata is then stored in the MySQL database. Note that attributes *FID-Real*, *FID-Imag*, and *Spectrum* are time-series data, hence the values of



https://lrs-tucker01.ist.psu.edu/szt5115/nmrseer/

NMRexSeer Demo

Query □ > Auto-wildcard

Advance Mode:

Advanced Search:

Query AlCl H8781 ☑ > Auto-wildcard

Operator: wash198

Instr. Vendor: BRUKER Nucleus: 27Al Pulse Sequence:

Observe Freq. Range 100.619169 to 849.7438002

Signal Avg. Range 8 to 100

Preacquisition Delay Range: to

Fig. 3. NMRexSeer search interface allows users to search for NMR experimental data with both *Simple Search* and *Advanced Search* modes.

these attributes are extracted and kept in separate CSV files.

NMRexSeer also generates *previews* as part of the metadata. The previews of each data package include the graphical representation of the real free induction decay (FID-Real), imaginary free induction decay (FID-Imag), and frequency domain spectrum (Spectrum), generated using GNU Plot tool [30]. These preview images could help users to visualize the NMR spectra before deciding to download the associated package.

B. Indexing and Searching

NMRexSeer indexer builds on *Sphinx Search Server* [7]. Sphinx has the ability to incrementally index records stored in a MySQL database. Sphinx also provides search APIs that allow multiple programming languages to interact and search the index.

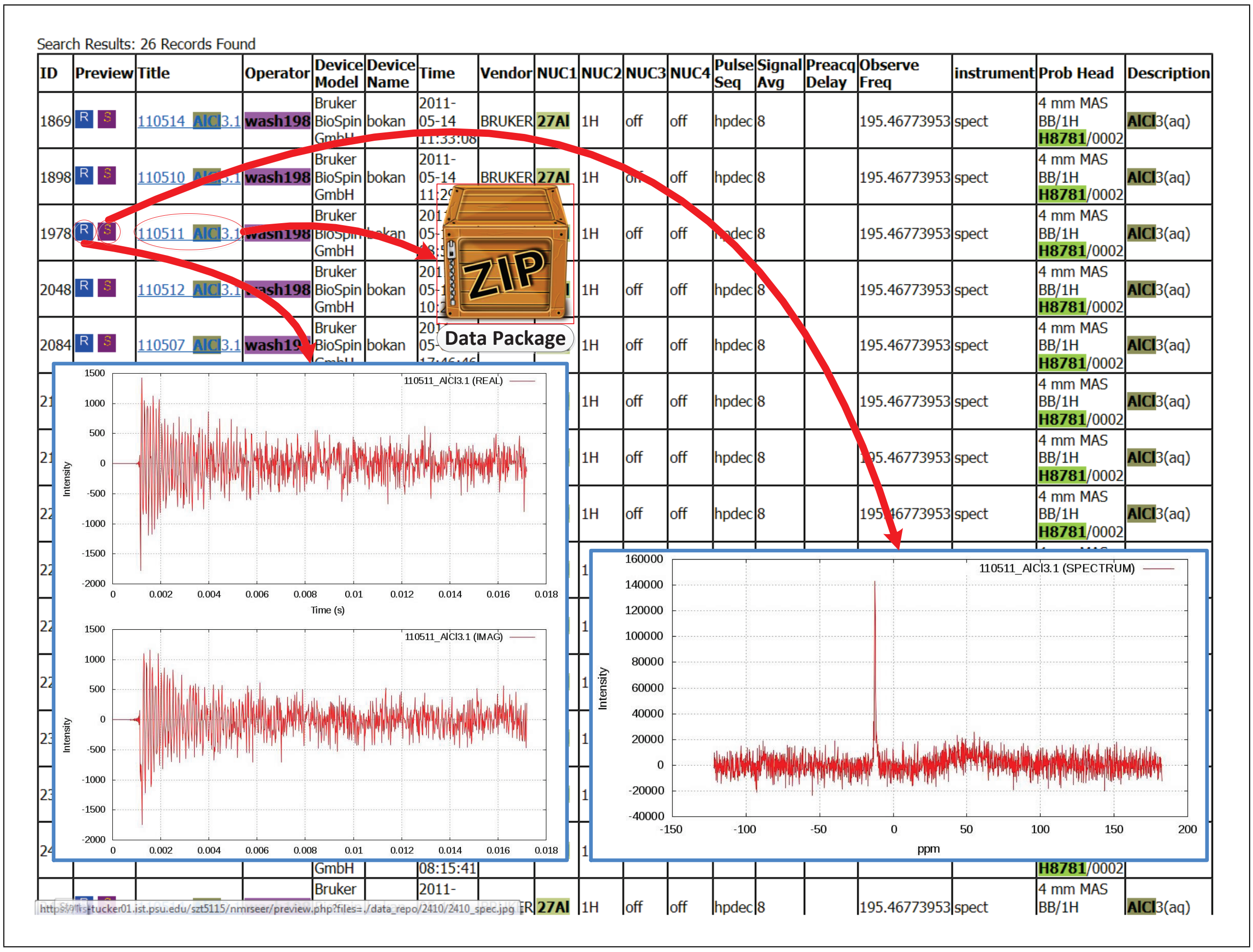
NMRexSeer search interface, illustrated in Figure 3, is implemented with PHP. The demo version is publicly available for test use[1]. The users can perform the search in two modes: *Simple Search* and *Advanced Search*. The Simple Search mode allows the user to input any free text query. If the "Auto-wildcard" is off, then NMRexSeer will retrieve records whose metadata contains an exact match to the query. However, if the "Auto-wildcard" mode is on, then partial match is performed. In the *Advanced Search* mode (activated by clicking the drop-down icon), in addition to the free text query, the user can further specify filtering criteria. Possible values of some filtering criteria such as *Instrument Vendor*, *Nucleus*, and *Pulse Sequence* are pre-computed and presented as dropdown menus for convenience. Internally, the query and filtering criteria are first transformed into an expression in Sphinx internal query language before further processing.

Figure 4 shows a sample search session using the search configurations in the *Advanced Mode* in Figure 3. In each matched record, the matched criteria are highlighted with different colors. The results are ranked using the combination of both phrase proximity and BM25. The user can examine the previews by clicking at either the *R* or *S* icon in the *Preview* field. The *R* icon leads to the visualized previews of both the real and imaginary part of the free induction decays (i.e. both *FID-Real* and *FID-Imag*). The *S* icon leads to the preview of the frequency domain spectrum. These visualized previews are regarded as useful by chemists at EMSL. Clicking the title of each search result activates downloading the corresponding data package.

IV. Conclusions and Future Work

We describe the architecture of *NMRexSeer*, an automated system that extracts nuclear magnetic resonance spectroscopy (NMR) specific metadata from NMR experimental data packages and makes them searchable. The proposed system is an extension to MyEMSL data repository for experimental data at

[1] https://lrs-tucker01.ist.psu.edu/szt5115/nmrseer/



Search Results: 26 Records Found

ID	Preview	Title	Operator	Device Model	Device Name	Time	Vendor	NUC1	NUC2	NUC3	NUC4	Pulse Seq	Signal Avg	Preacq Delay	Observe Freq	instrument	Prob Head	Description
1869	R S	110514_AlCl3.1	wash198	Bruker BioSpin GmbH	bokan	2011-05-14 11:33:08	BRUKER	27Al	1H	off	off	hpdec	8		195.46773953	spect	4 mm MAS BB/1H H8781/0002	AlCl3(aq)
1898	R S	110510_AlCl3.1	wash198	Bruker BioSpin GmbH	bokan	2011-05-14 11:29	BRUKER	27Al	1H	off	off	hpdec	8		195.46773953	spect	4 mm MAS BB/1H H8781/0002	AlCl3(aq)
1978	R S	110511_AlCl3.1	wash198	Bruker BioSpin GmbH	bokan	201 05-1 08:5			1H	off	off	hpdec	8		195.46773953	spect	4 mm MAS BB/1H H8781/0002	AlCl3(aq)
2048	R S	110512_AlCl3.1	wash198	Bruker BioSpin GmbH	bokan	201 05-1 10:2			1H	off	off	hpdec	8		195.46773953	spect	4 mm MAS BB/1H H8781/0002	AlCl3(aq)
2084	R S	110507_AlCl3.1	wash19	Bruker BioSpin	bokan	201 05-			1H	off	off	hpdec	8		195.46773953	spect	4 mm MAS BB/1H H8781/0002	AlCl3(aq)
21									1H	off	off	hpdec	8		195.46773953	spect	4 mm MAS BB/1H H8781/0002	AlCl3(aq)
21									1H	off	off	hpdec	8		5.46773953	spect	4 mm MAS BB/1H H8781/0002	AlCl3(aq)
22									1H	off	off	hpdec	8		195 46773953	spect	4 mm MAS BB/1H H8781/0002	AlCl3(aq)
22									1									
22									1									
22									1									
23									1									
23									1									
24				GmbH		08:15:41			1								H8781/0002	
	R S			Bruker		2011-	ER	27Al	1H	off	off	hpdec	8		195.46773953	spect	4 mm MAS BB/1H	AlCl3(aq)

https://...tucker01.ist.psu.edu/szt5115/nmrseer/preview.php?files=./data_repo/2410/2410_spec.jpg

Fig. 4. Example search session. Each search result has preview images of the NMR spectra and a pointer to download the data package.

the EMSL user facility, with the goal to build better metadata extraction for each experiment type to improve discovery and reuse of existing experimental data. Such a system could prove to not only facilitate remote data access, but also reduce wastage of resources and storage space. *NMRexSeer* was well received by chemists at EMSL and we were encouraged to move forward in this direction. Future work would involve building specific metadata extraction systems for other experiment types. These experiment-type specific metadata could then be integrated into a single, flexible metadata standard such as Investigation-Study-Assay (ISA)[31], Dubrin Core[32], and [33]. Such integration would enable cross-platform search and useful applications in data mining and knowledge discovery. Finally, we make the source code and sample data packages available for research purposes.

V. Acknowledgments

We gratefully acknowledge support from the Department of Energy, along with useful comments from Nancy Washton and Karl Mueller.

References

[1] Environmental molecular sciences laboratory. [Online]. Available: https://www.emsl.pnl.gov/emslweb/

[2] D. Cowley, "Myemsl & emslhub: Creating a flexible framework for scientific data sharing, discovery, and collaboration," Jun. 2013. [Online]. Available: http://www.institute.loni.org/lasigma/workshops/dataWorkshop2013/

[3] J. Keeler, *Understanding NMR spectroscopy*. John Wiley & Sons, 2013.

[4] Bruker: Nuclear magnetic resonance (nmr). [Online]. Available: https://www.bruker.com/products/mr/nmr.html

[5] Agilent technologies: Nuclear magnetic resonance (nmr). [Online]. Available: http://www.chem.agilent.com/en-US/products-services/Instruments-Systems/Nuclear-Magnetic-Resonance/Pages/default.aspx

[6] Emsl: Nmr and epr. [Online]. Available: http://www.emsl.pnl.gov/capabilities/nmr/

[7] Sphinx: Open source search server. [Online]. Available: http://sphinxsearch.com/

[8] Dress: A repository for solution nmr structures refined in explicit solvent. [Online]. Available: http://www.cmbi.ru.nl/dress/

[9] Nmrshiftdb2: Open nmr database on the web. [Online]. Available: http://nmrshiftdb.nmr.uni-koeln.de/

[10] Membrane proteins of known structure determined by nmr. [Online]. Available: http://www.drorlist.com/nmr/MPNMR.html

[11] Madison-qingdao metabolomics consortium database. [Online]. Available: http://mmcd.nmrfam.wisc.edu/

[12] P. Mitra, C. L. Giles, B. Sun, and Y. Liu, "Chemxseer: A digital library and data repository for chemical kinetics," in *Proceedings of the ACM First Workshop on CyberInfrastructure: Information Management*

in eScience, ser. CIMS '07. New York, NY, USA: ACM, 2007, pp. 7–10. [Online]. Available: http://doi.acm.org/10.1145/1317353.1317356

[13] N. Li, L. Zhu, P. Mitra, K. Mueller, E. Poweleit, and C. L. Giles, "orechem chemxseer: A semantic digital library for chemistry," in *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, ser. JCDL '10. New York, NY, USA: ACM, 2010, pp. 245–254. [Online]. Available: http://doi.acm.org/10.1145/1816123.1816160

[14] S. R. Choudhury, S. Tuarob, P. Mitra, L. Rokach, A. Kirk, S. Szep, D. Pellegrino, S. Jones, and C. L. Giles, "A figure search engine architecture for a chemistry digital library," in *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, ser. JCDL '13. New York, NY, USA: ACM, 2013, pp. 369–370. [Online]. Available: http://doi.acm.org/10.1145/2467696.2467757

[15] Z. Wu, J. Wu, M. Khabsa, K. Williams, H.-H. Chen, W. Huang, S. Tuarob, S. R. Choudhury, A. Ororbia, P. Mitra *et al.*, "Towards building a scholarly big data platform: Challenges, lessons and opportunities," in *Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference on*. IEEE, 2014, pp. 117–126.

[16] J. Wu, J. Killian, H. Yang, K. Williams, S. R. Choudhury, S. Tuarob, C. Caragea, and C. L. Giles, "Pdfmef: A multi-entity knowledge extraction framework for scholarly documents and semantic search," 2015.

[17] Y. Liu, K. Bai, P. Mitra, and C. L. Giles, "Tableseer: automatic table metadata extraction and searching in digital libraries," in *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2007, pp. 91–100.

[18] S. R. Choudhury, S. Tuarob, P. Mitra, L. Rokach, A. Kirk, S. Szep, D. Pellegrino, S. Jones, and C. L. Giles, "A figure search engine architecture for a chemistry digital library," in *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2013, pp. 369–370.

[19] S. Bhatia, S. Tuarob, P. Mitra, and C. L. Giles, "An algorithm search engine for software developers," in *Proceedings of the 3rd International Workshop on Search-Driven Development: Users, Infrastructure, Tools, and Evaluation*. ACM, 2011, pp. 13–16.

[20] S. Tuarob, P. Mitra, and C. L. Giles, "Improving algorithm search using the algorithm co-citation network," in *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*. ACM, 2012, pp. 277–280.

[21] ——, "A classification scheme for algorithm citation function in scholarly works," in *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital Libraries*. ACM, 2013, pp. 367–368.

[22] S. Tuarob, S. Bhatia, P. Mitra, and C. L. Giles, "Automatic detection of pseudocodes in scholarly documents using machine learning," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013, pp. 738–742.

[23] S. Tuarob, P. Mitra, and C. L. Giles, "Building a search engine for algorithms," *ACM SIGWEB Newsletter*, no. Winter, p. 5, 2014.

[24] ——, "A hybrid approach to discover semantic hierarchical sections in scholarly documents," in *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*. IEEE, 2015.

[25] S. Tuarob, L. C. Pouchard, P. Mitra, and C. L. Giles, "A generalized topic modeling approach for automatic document annotation," *International Journal on Digital Libraries*, vol. 16, no. 2, pp. 111–128, 2015.

[26] S. Tuarob, L. C. Pouchard, and C. L. Giles, "Automatic tag recommendation for metadata annotation using probabilistic topic modeling," in *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, ser. JCDL '13. New York, NY, USA: ACM, 2013, pp. 239–248. [Online]. Available: http://doi.acm.org/10.1145/2467696.2467706

[27] S. Tuarob, L. C. Pouchard, N. Noy, J. S. Horsburgh, and G. Palanisamy, "Onemercury: Towards automatic annotation of environmental science metadata," in *Proceedings of the 2nd International Workshop on Linked Science*, 2012.

[28] Onemercury: A search tool for scientific data. [Online]. Available: https://cn.dataone.org/onemercury/

[29] Mercury: Distributed metadata management, data discovery and access system. [Online]. Available: http://mercury.ornl.gov/

[30] gnuplot homepage. [Online]. Available: http://www.gnuplot.info/

[31] P. Rocca-Serra, M. Brandizi, E. Maguire, N. Sklyar, C. Taylor, K. Begley, D. Field, S. Harris, W. Hide, O. Hofmann, S. Neumann, P. Sterk, W. Tong, and S.-A. Sansone, "Isa software suite: supporting standards-compliant experimental annotation and enabling curation at the community level," *Bioinformatics*, vol. 26, no. 18, pp. 2354–2356, 2010. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/26/18/2354.abstract

[32] S. Weibel, J. Kunze, C. Lagoze, and M. Wolf, "Dublin core metadata for resource discovery," *Internet Engineering Task Force RFC*, vol. 2413, no. 222, p. 132, 1998.

[33] J. Bernard, T. Ruppert, M. Scherer, J. Kohlhammer, and T. Schreck, "Content-based layouts for exploratory metadata search in scientific research data," in *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, ser. JCDL '12. New York, NY, USA: ACM, 2012, pp. 139–148. [Online]. Available: http://doi.acm.org/10.1145/2232817.2232844