Rule-based Capture/Storage of Scientific Data from PDF Files and Export using a Generic Scientific Data Model

Stuart J. Chalk, Audrey Bartholomew, Bashar Baraz, John Turner
Department of Chemistry, University of North Florida
schalk@unf.edu

#ACSCINFDataSummit

UNF UNIVERSITY of NORTH FLORIDA.

CINF Paper 3 – 251st ACS Meeting Spring 2016

Outline

* Data, Data Everywhere
* Extracting Text from PDF Files
* Making Sense of Textual Data
* Rules and Rulesets
* Overcoming Incompatibilities
* A Generic Scientific Data Model
* Building the Website
* Functionality of the Website
* Next Steps
* Future Plans
* Conclusion

Data, Data Everywhere

* U.S. Gov. "… open and machine readable data…" (5/2013)
  * http://www.data.gov/
* NSF's Open Data Inventory (http://www.nsf.gov/data/)
  * Expand -- publish additional data assets in the inventory:
  * Enrich -- improve the discoverability, management, and reusability of agency data assets through metadata.
  * Open -- provide machine-readable and publically accessible agency data assets.
* Funding agencies require "Data Management Plans"
  * MPS Open Data Workshop - https://mpsopendata.crc.nd.edu/

Data, Data Everywhere (*but interoperable?*)

* Joint Declaration of Data Citation Principles
  Force 11 - https://www.force11.org/datacitation
  * Importance
  * Credit and Attribution
  * Unique Identification
  * Access
  * Persistence
  * Specificity and Verifiability
  * Interoperability and Flexibility

Extracting Text from PDF Files

* PDF Files are not meant to be vehicles for research data sharing – but sadly have become the default
* Force 11 grew out of an initial meeting at UC San Diego in 2011 – *Beyond the PDF* https://sites.google.com/site/beyondthepdf/
* PDF files store text in many places, and not in logical sequence as it appears on the page
* Scanned images can be converted to text using optical character recognition
* Many tools to get text out of PDF files

Extracting Text from PDF Files

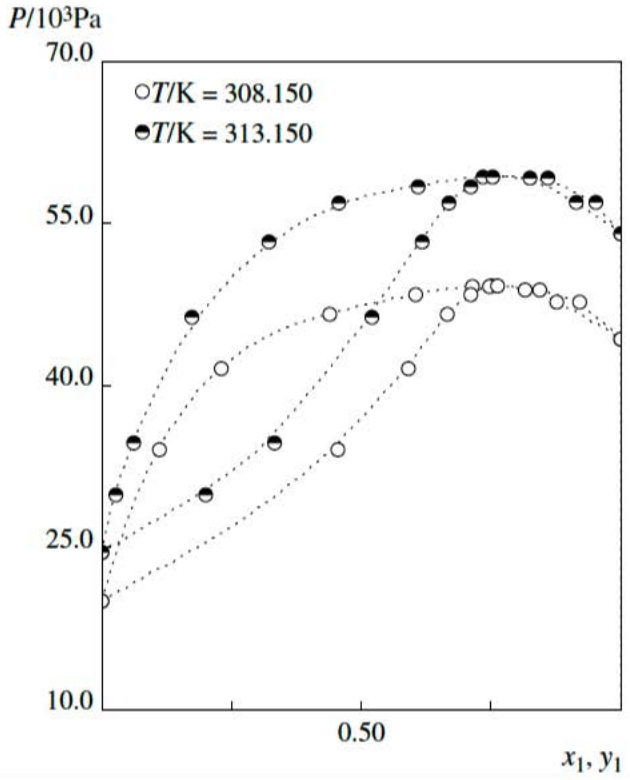
* Software tested
  * Adobe acrobat (plain text, accessible text, RTF, XML) Converted to the newest PDF version
  * lapdftext (Location Aware) https://github.com/GullyAPCBurns/lapdftext
  * pdftotext (part of) Xpdf http://www.foolabs.com/xpdf/home.html

Extracting Text from PDF Files

File Number: LB0062

Components: **1. $C_3H_6O_2$, Methyl ethanoate** [79-20-9]
2. C_6H_{12}, Cyclohexane [110-82-7]

T/K = 313.150		
x_1	$P/10^3$Pa	y_1
0.000000	24.6230	0.000000
0.028000	29.9440	0.201000
0.062000	34.7300	0.334000
0.174000	46.3030	0.521000
0.322000	53.3020	0.618000
0.457000	56.9420	0.670000
0.610000	58.4220	0.711000
0.735000	59.3420	0.754000
0.860000	59.2220	0.826000
0.952000	56.9690	0.915000
1.000000	54.0620	1.000000



Uncertainties: $\sigma(x_1) = 0.0010$; $\sigma(P)/10^3\text{Pa} = 0.030 - 0.070$; $\sigma(y_1) = 0.0010$ See SELF for the totality of data

Nagata, I.; Ohta, T.; Takahashi, T.; Gotoh, K. Thermodynamic properties of methyl acetate-benzene and methyl acetate-cyclohexane mixtures *J. Chem. Eng. Jpn.* **1973,** *6*, 129-134

Making Sense of Textual Data

63 H2 O (1) water 7732-18-5
C2 H4 O2 (2) acetic acid 64-19-7

T/C = 30.0 70W1
X2 0.00 0.03 0.05 0.08 0.13 0.22 0.35 0.51 0.69
ST 71.25 55.45 50.70 44.96 42.41 38.40 36.95 33.37 31.00
X2 0.85 1.00
ST 29.50 26.34
T/C = 18.0 02W1
W2 0.00 0.05 0.15 0.25 0.30 0.35 0.40 0.45 0.55
ST 74.16 61.17 51.55 46.89 45.02 43.55 42.45 41.05 38.73
W2 0.70 0.75 0.80 0.85 0.90 0.95 1.00
ST 36.37 34.90 33.51 32.12 30.84 29.32 27.56
T/C = 20.0 09G1
W2 0.0000 0.1060 0.2030 0.4390 0.5022 0.6780 0.7803 0.8760 0.9750
ST 75.23 56.63 48.30 35.27 35.01 31.45 28.26 26.98 24.97
T/C = 30.0 13M1
W2 0.000 0.0100 0.02475 0.05001 0.1001 0.1498 0.2009 0.3009 0.4011
ST 71.030 67.756 63.995 59.435 53.500 49.451 46.455 42.269 39.374
W2 0.4996 0.6005 0.6991 0.7988 0.9004 1.0000
ST 37.109 35.035 33.099 31.026 28.677 25.725
T/C = 20.0 47G2
W2 0.0000 0.1603 0.2940 0.4860 0.5246 0.5524 0.5690 0.6093 0.6380
ST 72.53 47.68 41.50 36.20 35.10 34.62 34.20 33.70 32.30
W2 0.7080 0.7540 0.7550 0.7800 0.8230 0.9268 0.9950
ST 31.70 30.73 30.73 29.01 28.27 25.40 22.42
T/C = 30.0 47G2
W2 0.0000 0.1603 0.2940 0.4860 0.5246 0.5524 0.5690 0.6093 0.6380
ST 71.03 45.90 39.40 32.70 32.30 31.40 31.30 31.00 29.30
W2 0.7080 0.7540 0.7550 0.7800 0.8230 0.9268 0.9950
ST 28.75 27.50 27.40 26.20 25.50 22.60 18.83

```
63 H2 O (1) water 7732-18-5
C2 H4 O2 (2) acetic acid 64-19-7

T/C = 30.0 70W1
X2 0.00 0.03 0.05 0.08 0.13 0.22 0.35 0.51 0.69
ST 71.25 55.45 50.70 44.96 42.41 38.40 36.95 33.37 31.00
X2 0.85 1.00
ST 29.50 26.34
T/C = 18.0 02W1
W2 0.00 0.05 0.15 0.25 0.30 0.35 0.40 0.45 0.55
ST 74.16 61.17 51.55 46.89 45.02 43.55 42.45 41.05 38.73
W2 0.70 0.75 0.80 0.85 0.90 0.95 1.00
ST 36.37 34.90 33.51 32.12 30.84 29.32 27.56
T/C = 20.0 09G1
W2 0.0000 0.1060 0.2030 0.4390 0.5022 0.6780 0.7803 0.8760 0.9750
ST 75.23 56.63 48.30 35.27 35.01 31.45 28.26 26.98 24.97
T/C = 30.0 13M1
W2 0.000 0.0100 0.02475 0.05001 0.1001 0.1498 0.2009 0.3009 0.4011
ST 71.030 67.756 63.995 59.435 53.500 49.451 46.455 42.269 39.374
W2 0.4996 0.6005 0.6991 0.7988 0.9004 1.0000
ST 37.109 35.035 33.099 31.026 28.677 25.725
T/C = 20.0 47G2
W2 0.0000 0.1603 0.2940 0.4860 0.5246 0.5524 0.5690 0.6093 0.6380
ST 72.53 47.68 41.50 36.20 35.10 34.62 34.20 33.70 32.30
W2 0.7080 0.7540 0.7550 0.7800 0.8230 0.9268 0.9950
ST 31.70 30.73 30.73 29.01 28.27 25.40 22.42
T/C = 30.0 47G2
W2 0.0000 0.1603 0.2940 0.4860 0.5246 0.5524 0.5690 0.6093 0.6380
ST 71.03 45.90 39.40 32.70 32.30 31.40 31.30 31.00 29.30
W2 0.7080 0.7540 0.7550 0.7800 0.8230 0.9268 0.9950
ST 28.75 27.50 27.40 26.20 25.50 22.60 18.83
```

Rules and Rulesets

* Process the text file one line at a time
  * 'Generic' rules to make decisions on a line
    * May be multiple rules for one line
    * Includes instructions to move down or stay on line
  * Sets of rules tailored to match the generic layout of the text
  * These were linked to datatype and publication

name	pattern	action	failure	valueName	errorText	required	example
GetID	^(\d+)	5	0	fileNum		0	1 Al Br3 aluminium bromide 7727-15-3
GetFormula	((\w+)+)	5	6	chemicalFormula	Chemical formula not found	1	1 Al Br3 aluminium bromide 7727-15-3
GetChemicalName	(([A-Za-z0-9-\)\(\]\[, \+\-]+)+)	5	6	chemicalName	Chemical name not found	0	1 Al Br3 aluminium bromide 7727-15-3
getCAS	(\d{2,6}-\d{2}-\d)	5	5	CAS		0	1 Al Br3 aluminium bromide 7727-15-3

More sophiticated GetFormula regex (([A-Z][a-zA-Z0-9]{0,2}\s)+)

Overcoming Incompatibilities

- The data we extract with the ruleset is saved in a json string
- But how should it be saved so that compatible with other data?
- Different properties
- Different units
- We need a generic data model to put the data in…


```
{
    "FileNum": "63",
    "ChemicalFormula": [
        "H2 O",
        "C2 H4 O2"
    ],
    "ChemicalName": [
        "water",
        "acetic acid"
    ],
    "CAS": [
        "7732-18-5",
        "64-19-7"
    ],
    "ParametersUnit": [["T/C"],["T/C"],["T/C"],["T/C"],["T/C"],["T/C"]],
    "Parameters": [["30.0"],["18.0"],["20.0"],["30.0"],["20.0"],["30.0"]],
    "References": ["70W1","02W1","09G1","13M1","47G2","47G2"],
    "DataUnits": ["X2","ST","W2","ST","W2","ST","W2","ST","W2","ST","W2","ST"],
    "Split": ["0","2","4","6","8","10","12","12"],
    "Data": [
        ["0.00","0.03","0.05","0.08","0.13","0.22","0.35","0.51","0.69","0.85","
        ["71.25","55.45","50.70","44.96","42.41","38.40","36.95","33.37","31.00"
        ["0.00","0.05","0.15","0.25","0.30","0.35","0.40","0.45","0.55","0.70","
        ["74.16","61.17","51.55","46.89","45.02","43.55","42.45","41.05","38.73"
        ["0.0000","0.1060","0.2030","0.4390","0.5022","0.6780","0.7803","0.8760"
        ["75.23","56.63","48.30","35.27","35.01","31.45","28.26","26.98","24.97"
        ["0.000","0.0100","0.02475","0.05001","0.1001","0.1498","0.2009","0.3009
        ["71.030","67.756","63.995","59.435","53.500","49.451","46.455","42.269"
        ["0.0000","0.1603","0.2940","0.4860","0.5246","0.5524","0.5690","0.6093"
        ["72.53","47.68","41.50","36.20","35.10","34.62","34.20","33.70","32.30"
        ["0.0000","0.1603","0.2940","0.4860","0.5246","0.5524","0.5690","0.6093"
        ["71.03","45.90","39.40","32.70","32.30","31.40","31.30","31.00","29.30"
    ]
}
```

A Generic Scientific Data Model

* What is data?
  A value of a measured property with/without unit
  … or a series of data
  ... or an observation (text)
* But on its own its not that useful...
* What's the context?
  * Under what conditions was the data obtained?
  * What were you testing?
  * What instrument did you use to take the measurement?
  * Who did the experiment?

A Generic Scientific Data Model

* How to store all this information?
  * Relational database – very structured, data has to fixed schema
  * Graph database – unstructured, flexible, non-standardized
* Intermediate solution needed
  * Must have some structure so it can be easily searched
  * Must be flexible to handle and type of data
* Approach
  * Use JSON for linked data (JSON-LD) as format to hold data
  * Provide a framework upon which you can hang the data and metadata

What is JSON for Linked Data?

* JavaScript Object Notation (JSON)
* JSON for Linked Data (JSON-LD)
  * W3C Recommentation (https://www.w3.org/TR/json-ld/)
  * Specification that allows data/metadata to be stored in JSON format but translated automatically to RDF (Resource Description Framework)
  * Provides a mechanism to add context to data/metadata stored in JSON
  * Note: Does not provide validation of the data/metadata
  * Can be validated using JSON Schema (http://json-schema.org)

What is JSON for Linked Data?

```
{
    "@context": {
        "name": "http://schema.org/name",
        "isAlive": "http://example.org/isAlive",
        "age": "http://example.org/age",
        "height": "http://schema.org/height",
        "@base": "http://www.unf.edu/chemistry/stuart_chalk.aspx"
    },
    "@id": "",
    "name": "Stuart Chalk",
    "isAlive": true,
    "age": 49,
    "height": 188.0
}
```

What is JSON for Linked Data?

```
<http://www.unf.edu/chemistry/stuart_chalk.aspx>
        <http://example.org/age>
                "49"^^<http://www.w3.org/2001/XMLSchema#integer> .

<http://www.unf.edu/chemistry/stuart_chalk.aspx>
        <http://example.org/isAlive>
                "true"^^<http://www.w3.org/2001/XMLSchema#boolean> .

<http://www.unf.edu/chemistry/stuart_chalk.aspx>
        <http://schema.org/height>
                "188"^^<http://www.w3.org/2001/XMLSchema#integer> .

<http://www.unf.edu/chemistry/stuart_chalk.aspx>
        <http://schema.org/name>
                "Stuart Chalk" .
```

http://json-ld.org/playground/

Building the Website

* CakePHP 2.7 PHP Framework (http://cakephp.org)
* Apache 2.4 webserver (http://httpd.apache.org)
* PHP 5.6 (http://www.php.org)
* MySQL 5.5 (http://www.mysql.com)
* jQuery (JavaScript) (http://jquery.com)
* Bootstrap (http://getbootstrap.com/)

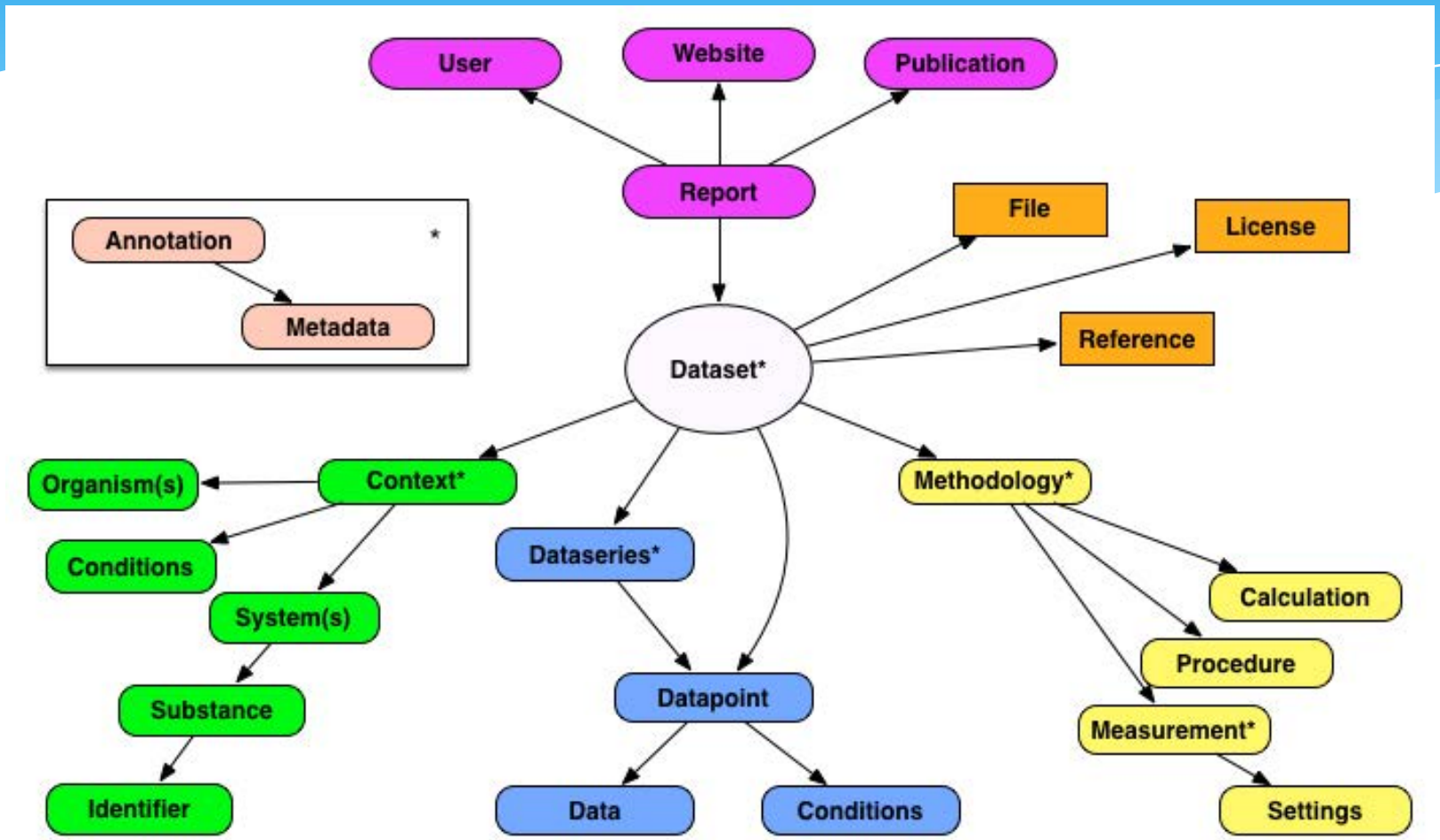
* PHPStorm (https://www.jetbrains.com)
* GitHub (https://github.com)

Functionality of the Website

* Upload PDF
* Assign Ruleset
* Extract data
* Perform QC
* Ingest Data
* View “dataset”

* Mass Process – many files from one publication
* Map the Process – see what happened

Database Schema



Live Demo

Next Steps

* Formalize the system to
  * Organize the processing
  * Capture any processing current done in scripts
  * Publish a paper
* Ingest data in SDM format into triplestore and perform SPARQL queries on aggregated data
* Aggregate data of the same type on the website and view, organize, plot together
* Provide option to capture and present equation based results

Take Home

* ~27,500 PDF pages processed
* ~530,000 pieces of physical property data extracted
* Data can be extracted from PDF files as long as the
  * Text is cleaned up
  * Context is encoded in the process of extraction
* Data homogeneity can be achieved by using a generic scientific data model
* More work needed to create robust system to handle data in any "well structured" format

Questions?

Thanks to Michael Klinge and Stefan Scherer

* schalk@unf.edu
* Phone: 904-620-5311
* Skype: stuartchalk
* LinkedIn/Slidehare: https://www.linkedin.com/in/stuchalk
* ORCID: http://orcid.org/0000-0002-0703-7776
* ResearcherID: http://www.researcherid.com/rid/D-8577-2013

ChemExtractor: Semantic Data from PDF Files

* Text Extraction of 11 Landolt-Börnstein Volumes
* Regulat Expression (Regex) fragments used to develop rules to detect specific types of data
* PHP Scripts used to apply 'Rulesets' to text
* Extracted data and metadata captured in JSON
* JSON data added to MySQL database to support searching and export in a generic scientific data model (JSON-LD)
* Website developed to search and display data
* ~27,500 PDF pages processed
* ~530,000 pieces of physical property data extracted