Development of File Format and Database Infrastructure for High Explosive Reference Data

Charles B. Kiyanda, Michelle M. Boyce and Hoi Dick Ng
Department of Mechanical and Industrial Engineering
Concordia University
Montreal, QC, Canada

1 Introduction

Underpinning the work of model development, in any discipline, is the ability to compare model and calculation results to experimental data. Examples of interest include: determining equation of state parameters for high explosives (HE) using experimental shock Hugoniot data [1, 2], determining reaction rate parameters for detonation front curvature measurements [3], etc. As such, the ability to use previously published data is of prime importance. Usage can be broken down into different tasks e.g.: searching past experimental data, evaluating data relevance to a particular need, and integrating data into a workflow either as data input, e.g., to drive a model optimization, or as a benchmark to gauge the effectiveness of a given model. These different activities require a convenient database to hold the experimental data, a means for collaborators to add experimental data into the databse (import), and a means of transforming the data to a more readily useful format (export). Poorly executed databases can sometimes lead to significant errors. See, e.g., this example from bioinformatics [4].

There is currently no solution for the convenient archival, exploration, and access of high explosives relevant experimental and/or reference data. There is a database centered on gaseous detonation data which has been developed and is hosted by the Graduate Aeronautics Laboratory at Caltech [5]. That solution can be considered to have been ahead of its time when developed and it has proven to be extremely useful to many researchers over the years, including some of the authors. However, it lacks the means for outside collaborators to conveniently add information to the database. The web interface also seems to be purely driven by static html, which most likely provides extra work for the maintainers. All the underlying data sets are accessible as plain text files, though they contain only minimal metadata (in the form of column headers). Most notably, the text file datasets lack units (which are assumed throughout the database) and citation information. While that information is on the database website, the fact that it is not grouped within a single file makes it more likely that information loss occurs as data files are transferred from researcher to researcher. In such a case, one then needs to refer back to the original source, in this case the web database, to recover the whole information. Some information was reported to the authors that a more "dynamic database" design was attempted or built, though subsequently retired, at Caltech. The reasons behind preserving only the current, more static design, are unknown to us.

Our objective is to build a complete and convenient toolchain accessible to the condensed explosive community to collect, explore, share, and reuse experimental data. We are reporting here on a first iteration of a workflow that includes a metadata-rich file format and a web exploration tool.

2 Infrastructure Components

The suite of computing tools that we present allows the archival, exploration, and access to reference data. The workflow toolchain is divided into three components: a metadata-rich file format, a database to facilitate data searches and data exploration, and a web-based gui to easily acces the database, visualize the data, and export to other formats. This workflow is schematized in Fig 1. The reference data is converted from its original source, a book, a journal paper, etc. to our metadata-rich file format. These files are parsed and included in a database which is accessed by a web interface for direct visualization or to produce files, possibly in a simpler format, that can be plotted with any tool.

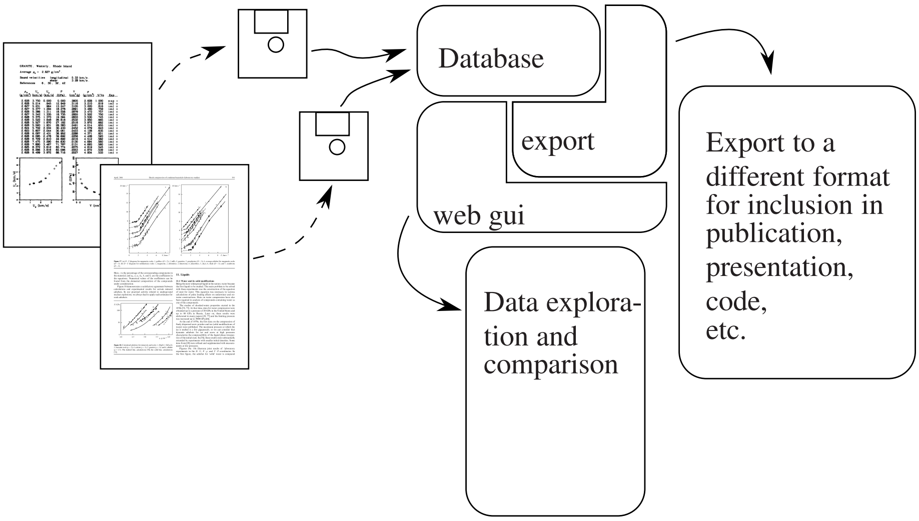

Database
export
web gui
Export to a different format for inclusion in publication, presentation, code, etc.
Data exploration and comparison

Figure 1: Workflow from published data to end use.

2.1 Metadata Rich File Format

At the core of the system is a metadata-rich file format that holds not only the actual experimental data, but also its units and other vital information that currently tends to get lost as researchers assemble local troves of reference data. Such information includes: data origin and citation information, experimental methods used to originally produce the data, material tested, material characteristics, precision of data, etc. Including this metadata along with the actual experimental data makes these files the basic unit of the system. Each file is a machine-readable version of the published data and researchers can freely copy and distribute those files when sharing data with no fear of information loss or degradation in the process.

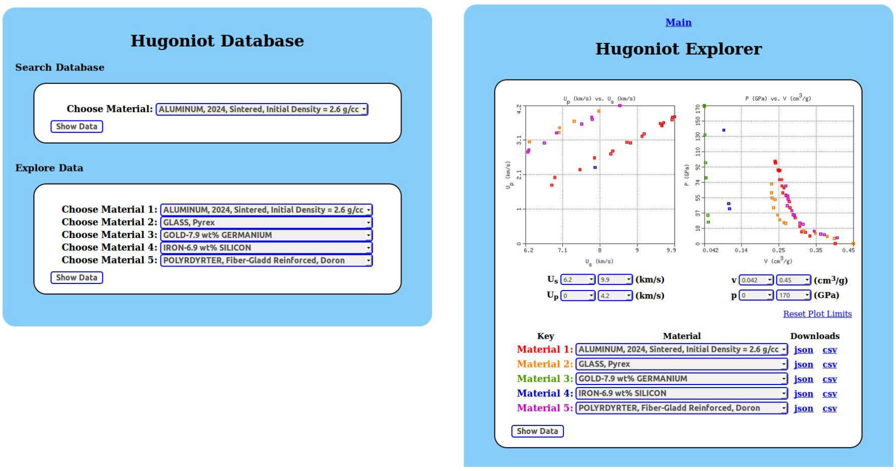
Since many datasets, especially when it comes to high explosives, can be very short, individual researchers may be justifiably tempted to simply read the file data when comparing data from different sources. As such, it is important that the adopted file format be also human-readable. We have thus devised a json based file format which aims at a balance between human readability and ease of parsability. Json being an established, de-facto standard [6], there are several tools to automatically import such files using most programming languages.

Json is a text based format and a representation of the javascript object structure. The format of a data file is best described by an example, here using data from [7] for pressed graphite. The same information was formatted, by hand, as table 1. All the information, including the citation information, is contained in the data file.

```
{"dataset name": "CARBON , graphite , pressed,
                  Initial density =  2.13 g/cc",
 "material":{
     "names" : ["carbon","graphite","pressed graphite"],
     "composition": {"C":1}
     },
 "data type" : "hugoniot",
 "reference":{ "type" : "book",
               "title":"LASL shock Hugoniot data",
               "authors": [{"firstname":"Stanley",
                            "middlename":"P.",
                            "lastname": "Marsh"}],
               "volume":5,
               "year":1980,
               "publisher":"University of California Press"
             },
 "data":{
   "variables":[
     {"name":"initial density"    ,"repr":"rho_0","units":"g/cc"},
     {"name":"initial temperature","repr":"T_0"  ,"units":"K"},
     {"name":"initial pressure"   ,"repr":"P_0"  ,"units":"GPa"},
     {"name":"shock velocity"     ,"repr":"U_s"  ,"units":"km/s"},
     {"name":"particle velocity"  ,"repr":"U_p"  ,"units":"km/s"},
     {"name":"pressure"           ,"repr":"P"    ,"units":"GPa"},
     {"name":"specific volume"    ,"repr":"v"    ,"units":"cc/g"},
     {"name":"density"            ,"repr":"rho"  ,"units":"g/cc"},
     {"name":"compression ratio"  ,"repr":"v/v_0","units":""}
    ],
   "contents":[
     {"point":[2.113, 273.15, 0.0, 5.235, 1.026,
               11.349, 0.3805, 2.628, 0.804], "comments":"im1" },
     {"point":[2.123, 273.15, 0.0, 6.013, 1.380,
               17.617, 0.3629, 2.755, 0.770], "comments":"im1" },
     {"point":[2.123, 273.15, 0.0, 6.320, 1.972,
               26.459, 0.3241, 3.086, 0.688], "comments":"im1" },
     {"point":[2.143, 273.15, 0.0, 6.551, 2.607,
               36.599, 0.2809, 3.560, 0.602], "comments":"im1" },
     {"point":[2.141, 273.15, 0.0, 6.704, 2.779,
               39.888, 0.2735, 3.657, 0.585], "comments":"im1" },
     {"point":[2.146, 273.15, 0.0, 7.960, 3.370,
               57.567, 0.2687, 3.722, 0.577], "comments":"im1" },
     {"point":[2.142, 273.15, 0.0, 8.762, 3.748,
               70.343, 0.2672, 3.743, 0.572], "comments":"im1" },
     {"point":[2.134, 273.15, 0.0, 8.836, 3.801,
               71.672, 0.2670, 3.745, 0.570], "comments":"im1" },
     {"point":[2.135, 273.15, 0.0, 9.208, 3.948,
               77.614, 0.2676, 3.737, 0.571], "comments":"im1" },
     {"point":[2.136, 273.15, 0.0, 9.627, 4.138,
               85.091, 0.2669, 3.746, 0.570], "comments":"im1" },
     {"point":[2.136, 273.15, 0.0, 9.566, 4.290,
               87.657, 0.2582, 3.873, 0.552], "comments":"im1" }
   ],
   "comments":"References 5,6,14\nAverage density = 2.134 g/cc"
   }
}
```

ρ_0[g/cc]	T_0[K]	P_0[GPa]	U_s[km/s]	U_p[km/s]	P[GPa]	v[cc/g]	ρ[g/cc]	v/v_0
2.113	273.15	0.0	5.235	1.026	11.349	0.3805	2.628	0.804
2.123	273.15	0.0	6.013	1.380	17.617	0.3629	2.755	0.770
2.123	273.15	0.0	6.320	1.972	26.459	0.3241	3.086	0.688
2.143	273.15	0.0	6.551	2.607	36.599	0.2809	3.560	0.602
2.141	273.15	0.0	6.704	2.779	39.888	0.2735	3.657	0.585
2.146	273.15	0.0	7.960	3.370	57.567	0.2687	3.722	0.577
2.142	273.15	0.0	8.762	3.748	70.343	0.2672	3.743	0.572
2.134	273.15	0.0	8.836	3.801	71.672	0.2670	3.745	0.570
2.135	273.15	0.0	9.208	3.948	77.614	0.2676	3.737	0.571
2.136	273.15	0.0	9.627	4.138	85.091	0.2669	3.746	0.570
2.136	273.15	0.0	9.566	4.290	87.657	0.2582	3.873	0.552

Table 1: Shock Hugoniot data for pressed graphite with an initial density of ρ_0 = 2.13 g/cc from [7].



Choose Material: ALUMINUM, 2024, Sintered, Initial Density = 2.6 g/cc

Choose Material 1: ALUMINUM, 2024, Sintered, Initial Density = 2.6 g/cc
Choose Material 2: GLASS, Pyrex
Choose Material 3: GOLD-7.9 wt% GERMANIUM
Choose Material 4: IRON-6.9 wt% SILICON
Choose Material 5: POLYRDYRTER, Fiber-Gladd Reinforced, Doron

U_s 6.2 9.9 (km/s)
U_p 0 4.2 (km/s)
v 0.042 0.45 (cm^3/g)
p 0 170 (GPa)

Material 1: ALUMINUM, 2024, Sintered, Initial Density = 2.6 g/cc
Material 2: GLASS, Pyrex
Material 3: GOLD-7.9 wt% GERMANIUM
Material 4: IRON-6.9 wt% SILICON
Material 5: POLYRDYRTER, Fiber-Gladd Reinforced, Doron

(a) (b)

Figure 2: Screenshot of the current web interface (a) main page and (b) sample data visualization.

Other file formats were considered beyond JSON including those described in table 2. An emphasis was put on a file format that is human-readable, ruling out binary file formats such as HDF5. Computational efficiency in treating datasets was not a very important factor as the datasets of interest to the community are typically small. We considered it essential to have access to standard parsing tools that would make it easy to enforce compliance of our file format. This requirement eliminated pure text based file formats, like CSV. While any software could read a file, a parser would have to be written in every language to enforce our desired file format. Using JSON or XML has the benefit that we can leverage established parsers, which are plentiful. Finally, verbosity was considered. On inspection, an XML based format visibly contained more words, simply, making it less human readable.

Format name	Human Readability	Std Parsing Tools	Computationally Efficient	Verbosity
text/CSV	Highest	None	No	Low
XML	Average/low	Yes	Low	High
HDF5	No (binary)	Yes	High	Comperssed
JSON	Average/high	Yes	Low	Average

Table 2: Alternative formats considered with their advantages and drawbacks.

2.2 Database and Web Access

Given a collection of such data files, we built a database that, once the metadata from the different datasets has been parsed, can be used more effectively to search and cross-reference the available data. This database is not a strictly necessary component, but it has the possible advantage of accelerating the exploration of the data. The notion of “importing” the data is therefore not the action of adding the information of a particular data file to the database, but rather consists in the conversion of the reference data from its original form to the metadata-rich file format. The database is merely a computational tool to facilitate the exploration of the datasets. The web interface should allow a user to:

1. query the database, thus returning the datasets of interest,
2. plot the data and explore that data dynamically,
3. possibly import new data through this web interface (i.e., the user could type, in a webpage, new data to include to the collection along with the relevant metadata, and a server then produces an appropriately formatted file and includes it in the database),
4. export the data to simpler data formats (csv, excel, etc.) that other plotting programs can understand.

Our current iteration of the system, a screenshot of which is shown in Fig 2, satisfies points 2 and 4. Point 1, database queries, is currently limited to listing all the available datasets. Point 3, data import through a web interface, is currently unimplemented. Right now, a user of the system needs to manually convert a dataset into the appropriate file format. The inclusion of that data file in the database is done manually on the server. The exploration of the data is also not as dynamic as it could be. Currently, when a request is made for a plot, a script on the server plots the requested data using gnuplot and generates a static image. This static image is then shown to the user. We plan, in the future, to have a more dynamic, possibly client-side application, that generates a plot that can be manipulated in real-time by the user. This is possible using e.g. javascript frameworks, notably, D3.js [8] or others. The current implementation is made available through the Concordia University Combustion and Energy Research Group at `http://users.encs.concordia.ca/~hoing/database.html`.

4 Future Direction

The nature of this project is perfectly suited to working with open-source tools and licensing the tools we produce as open-source code. While our focus is on experimental data relevant to high explosives research, a general enough file format could be reused in any research field. The time-consuming part of this project is to convert, i.e. digitize, historical and published data so it can be expressed into this file format. This must normally be done manually though we can use scanning and optical character recognition to accelerate the process. Adding an "import" option to the web gui serves, in essence, to distribute the workload so that any researcher freely using this database can submit data of interest not yet in the collection. Access to the web tools and database should be free and open to encourage the participation of any researcher in the field that wants to collaborate. There are no plans to build into the toolchain a moderation step, i.e. a process by which chosen or vetted individuals would curate the database (beyond cleaning obvious mistakes or vandalism). While there can be concerns about keeping both the quality of the database (e.g. ensuring that the data in the database accurately reflects the data collected by the researchers) as well as the quality of the data itself (i.e. that imprecise, poor scientific data is not included), we believe this concern is not a major one. There are examples of user-curated, non-centrally moderated databases that result in a high quality data source. An example is the OpenStreetMap project [9], that aims at creating open map data. It is possible that, to be effective, community contributed and curated datasets require a change tracking infrastructure. Such a solution is implementable on top of our system. We believe our approach, analogous to the spirit of the free software movement, should mesh naturally with the social conventions of scientific fields.

References

[1] Holmes NC, Moriarty JA, Gathers GR, and Nellis WJ. The equation of state of platinum to 660 GPa (6.6 Mbar). *Journal of Applied Physics*, 66(7):2962–2967, 1989.

[2] Heinz DL and Jeanloz R. The equation of state of the gold calibration standard. *Journal of Applied Physics*, 55(4):885–893, 1984.

[3] Bdzil JB. Steady-state two-dimensional detonation. *Journal of Fluid Mechanics*, 108:195–226, 1981.

[4] Zeeberg B, Riss J, Kane D, Bussey K, Uchio E, Linehan WM, Barrett JC, and Weinstein J. Mistaken identifiers: Gene name errors can be introduced inadvertently when using excel in bioinformatics. *BMC Bioinformatics*, 5(1):80, 2004.

[5] Kaneshige M and Shepherd JE. Detonation database. technical report fm97-8, july 1997. Available at http://www2.galcit.caltech.edu/detn_db/html/db.html.

[6] ECMA-404, the JSON data interchange format, 2013. Available at http://ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf.

[7] Marsh SP. *LASL Shock Hugoniot Data*, volume 5. Univ of California Press, 1980.

[8] D3.js website. Available at http://d3js.org/.

[9] Openstreetmap website. Available at http://www.openstreetmap.org/.