A Collaborative Data Library for Testing Prognostic Models

Joanna Sikorska[1], Melinda Hodkiewicz[2], Ashwin D'Cruz[3], Lachlan Astfalck[4], and Adrian Keating[5]

[1]*CASWA, 2/33 Horus Bend, Bibra Lake, Perth, Western Australia, 6163, Australia (jo@caswa.com)*

[2] *The University of Western Australia, School of Mechanical and Chemical Engineering, Mail Bag M050, Western Australia, 6009, Australia, (melinda.hodkiewicz@uwa.edu.au)*

[3] *The University of Western Australia, School of Mechanical and Chemical Engineering, Mail Bag M050, Western Australia, 6009, Australia, (ashwin.dcruz@uwa.edu.au)*

[4] *The University of Western Australia, School of Mechanical and Chemical Engineering, Mail Bag M050, Western Australia, 6009, Australia, (Lachlan.astfalk@uwa.edu.au)*

[5] *The University of Western Australia, School of Mechanical and Chemical Engineering, Mail Bag M050, Western Australia, 6009, Australia, (adrian.keating@uwa.edu.au)*

ABSTRACT

A web-based data management system for use by researchers and industry around the world to access suitable datasets for testing prognostic models is developed. The value of the project is in the provision of, and access to, real-world data for asset failure prediction work. In practice, it is difficult for researchers to obtain data from industrial equipment. Industry datasets are rarely shared and hardly ever published. When such data is made available, very little meta-data about the underlying asset is provided. This restricts the number and type of models that can be applied.

The solution is a data management system for three groups: researchers needing datasets, industry and academics with datasets. This paper identifies the data being sought, the system requirements and architecture, and discusses how the design is being implemented using an Agile development approach. Crucially, meta-data is stored in the database and accessed using a secure web-based front-end so as to maximize the available information, whilst obfuscating any corporate-sensitive material. The success of this prognostics data library depends on the support of the prognostic community to contribute and use the data; similar projects have been successful in the Machine Learning and Big Data communities.

1. INTRODUCTION

It is crucial for industry to manage the business risks associated with failures of its operational assets. Current practice is able to identify when many assets are starting to degrade and often, can even identify the way that the asset is failing. However, determining how long these assets can remain in service is still largely guesswork and based on the experience of personnel familiar with the equipment. Unfortunately, as equipment reliability increases (resulting in less frequent failures) this experience is becoming increasingly harder to acquire.

Prior work (Heng, Zhang, Tan, & Mathew, 2009; Jardine, Lin, & Banjevic, 2006; Sikorska, Hodkiewicz, & Ma, 2011; Vachtsevanos, Lewis, Roemer, Hess, & Wu, 2006) concluded that the ongoing implementation of most types of prognostic models is hampered by insufficient data for testing the proposed models. Even when some data has been collected to test a particular model, it is often inadequate to truly test the models' reliability and robustness in a real world context. Collected data usually relates to the following: (a) a single asset type operating under a limited set of operating conditions, (b) special laboratory rigs that can ensure a particular failure model, (c) simulations, or (d) diagnostic data. In the last case, the diagnostic data that is utilized for prognostics is rarely evaluated to determine whether the measured parameters are the most appropriate for predicting asset failure.

Further hampering the advancement of engineering prognostics, is that most research has been done on a very limited number of datasets that have been used repeatedly. This may seem advantageous because it allows new models to be compared easily with predecessors, however it also means that models have had very limited exposure to the variety of conditions and faults that occur in the real world. In order to be robust and useful predictors of failure, new models should be tested against a wider range of datasets that reflect the range of conditions they will be ultimately exposed to.

Unfortunately, getting new datasets is difficult. Although much cheaper that in decades past, data acquisition is still costly and time consuming. Opportunities for academics to gain access to real world plants for the purposes of collecting meaningful prognostic data are rare. Ironically, industry often has more data than it knows what to do with (albeit often imperfect) (Hodkiewicz, Kelly, Sikorska, & Gouws, 2006; Lee & Strong, 2004); data that is seldom shared and hardly ever published. Concerns about commercial confidentiality often prevent collaboration with outside parties.

Thus, for the field of engineering prognostics to progress more real-world data or realistic laboratory data, must be obtained. The data should be managed in such a way so as to maximize its utilization, whilst simultaneously maintaining any required privacy and commercial secrecy interests of its providers.

Our research is developing a data management system to pool and maximize the utilization of, suitable data for testing prognostic models. Data repositories to enhance collaboration exist already in the machine learning and big data communities but there are few asset health prognostics data sets. Individual prognostic research groups have also made prognostics data available but the number of data sets is small and they are mostly derived from simulations or experimental test rigs. Ultimately, the aim of this Prognostics Data Library (PDL) is to collate and enable sharing of high quality datasets across the entire prognostics community.

At the time of writing, development of the PDL is still underway. The project is using an Agile Development approach that requires ongoing involvement from stakeholders. Thus, we present our work to date in order to seek engagement with the wider prognostics community and ensure that the final outcome will best meet the needs of all potential users.

2. BACKGROUND AND JUSTIFICATION

2.1. Review of existing data sets and their availability

A number of sites have been created for sharing datasets and making them available to third party researchers. Two of the best known are the MIT Big Data Initiative (MIT, 2016) and the UC Irvine Machine Learning Repository (UCI) (Lichman, 2013). The UCI Repository contains more than 300 data sets mainly as .csv files. It is widely used by the machine learning and statistical communities to test new models. The most popular data set on this site, the Iris data set, has received almost a million hits and is cited by over 100 publications. We have found only one data set on this site that is the relevant to prognostics. It contains simulated data relating to gas turbine engines. Within the prognostic community NASA has been a leader in generating and sharing prognostics data through the Prognostics Data Repository (Goebel, 2015) and NASA DASHlink (NASA, 2016). There are 11 data sets in the repository and these have been used extensively by researchers around the world. The data sets are a mix of simulation, laboratory and field data, mainly in the form of time series. There are also a smaller number of data sets available, three on test rig data on turbine, pumps and bearings in the Acoustics and Vibration Database (Acoustics and Vibration Database, 2013), one data set from MFPT for a bearing test rig (Bechhoefer, 2013), and one from MaHeMM on a rail turnout system (MaHeMM, 2009). The CALCE laboratory is also a generator of battery and electronics failure data sets (CALCE, 2012). There is also a facility to upload data on the ResearchGate site but it is not widely used for engineering publications. In all cases it is a requirement that when datasets are used that a specific citation to acknowledge the original author is made.

Data challenges have also been a useful way of developing data sets. Every year at the Prognostics Health Management conference a new dataset is presented and participants must estimate the remaining useful life of the item. This data is available via various websites and is usually posted by the institution that developed the dataset. It includes the following: jet engine degradation simulation (NASA, 2008), gearbox health assessment (test rig data) (PHM Society, 2009), milling machine cutters RUL estimation(PHM Society, 2010), anemometer fault detection (PHM Society, 2011). In addition to these competitions the IEEE PHM Data Challenge has produced two data sets (IEEE, 2012, 2014).

As expected, there is no consistency in the format of currently available datasets. Some datasets separate the available data into modelling and testing files, others split different failure events into individual files, whilst others aggregate all data into one file. In most cases, datasets are provided in CSV format or as Matlab '.mat' files. ('Mat' files use a binary file format that requires to be parsed using a mathematics or programming language such as Matlab, R, Python, or Labview.) The amount and quality of meta-data about the underlying asset to which the failure relates differs greatly between datasets. In some cases commercial confidentiality is quoted as the reason not to provide any asset data. When provided, asset or file meta-data is usually presented as a separate text or pdf file that is not be easily decoded by a computer program building a model. Consequently, trialling

a new model on more than one dataset (even if sourced from the same website) requires computer programs to be partly re-coded for each trial (namely the lines associated with accessing, pre-processing and parsing data).

Despite these challenges there is widespread enthusiasm and support from stakeholders, both data providers and data users, for more work in this area. Stakeholders recognize the need to advance the development and use of algorithms and modelling through a collaborative, sharing approach.

2.2. Classes of Data Used in Prognostic Modelling

For robust prognostic modeling, data is required to adequately describe how an asset 'behaves' on its path from an incipient fault to final, catastrophic failure. Increasingly we are finding that failure behavior cannot be adequately modelled by physics of failure models or even by models based on internal covariates (such as sensor data) alone. In industry failure behavior is often affected by mission profile, operating conditions, the external environment and maintenance interventions (Jouin, Gouriveau, Hissel, Pera, & Zerhouni, 2016). Although in theory, physics-of-failure models should be able to take into account all of these input types, in practice either failure mechanisms are not perfectly understood under the range of actual input conditions, or there is insufficient accurate and multivariate data to verify and validate these models. Consequently, physical models tend to be limited to applications for well-understood faults in simple systems and/or by users with established diagnostic systems and predictive maintenance programs (Sikorska et al., 2011).

There is a growing body of work examining the issue of quantification and uncertainty in prognostics. There is a recent review of developments by Sankararaman (2015). While much of the work on uncertainty is currently focused on theoretical developments eventually these ideas will need to be tested on real examples. Data sets that include factors such as changes to mission profile, which may affect estimates of uncertainty, will be needed.

We have identified a number of categories of data that are useful for prognostic modelling. Wherever possible they are grouped using commonly used statistical modelling terms:

1. **Internal covariate data**: Data that is only available for the duration that either: (a) the underlying failure mode is progressing, or (b) while the asset suffering from the failure is operating.
    a. ***Direct failure mode data***: This dynamic (temporal) data is obtained from sensors that measure the actual failure occurring (e.g. acoustic emission from cracks or wear, volume of wear debris, actual crack length). This type of data is rare, hard to detect, will only be available once a failure mode has commenced, and is usually only available whilst that failure is actively occurring. Regular, or continuous monitoring of the exact area in which the fault is occurring, is usually needed. However, for single mode failures, it may also be the easiest type of data from which to develop simple remaining life predictors.
    b. ***Indirect failure mode data:*** This second type of dynamic (temporal) data relates to parameters that describe how the asset is responding to a failure mode that has already commenced (e.g. pump bearing vibration, exhaust gas temperature change, pipe wall thickness remaining). Collective experience has identified that similar assets react to specific failure modes in similar ways. As this is the most common type of data collected for diagnostics, it also the most used data type in prognostic models. This type of data is only available once failure has commenced and is sufficiently large to cause the asset to respond accordingly; it is also only available for the duration of the response condition. Multiple failure modes often cause the same measurable response; this is useful in that multiple failure modes may be detected using the same sensor, but conversely, they may make it difficult to identify which failure mode is underway. Coincident failure modes may be even more problematic to diagnose. Therefore, models based on indirect failure mode parameters need to be multi-variate so as to correctly identify which of several different outcomes is likely to eventuate.
2. **Situational or external covariates**: Dynamic (temporal) data that relates to parameters that are independent of the asset and its failure mode (i.e. external). These parameters are useful for prognostics because they describe conditions that may determine which failure mode is initiated, or how quickly it develops thereafter (e.g. weather conditions, operating profile, fluid characteristics). Unlike internal covariates, these data are available prior to, during and after failure. Unfortunately, this type of data is typically stored in disparate systems, is coarser than failure mode data, and is often overlooked when building prognostic models as it doesn't relate to the failure progression itself.
3. **Asset meta-data**: Static (unchanging) data describing what asset is being modeled and information about it operating environment. The design of the asset and its usual operating conditions (e.g. location, pumpage, type of site) may have a significant effect on how it will respond to conditions that initiate and promote failure. It is classified separately, as this constitutes data that does not change during the asset's life. It is rarely used by researchers in their prognostic models, probably because the models being developed are only confined to a limited number of assets. As prognostic models become more generic, or are incorporated into larger decision

making systems, this type of data will become increasingly important.

Some models use data in its raw format (direct from sensors) whilst other models require pre-processing to extract pertinent signal features. In the latter case, it is imperative to know how the data was extracted from the raw signals, so that the process can be replicated.

When using a third party file it is also crucial to know how to read the file and interpret its contents. We will hereby refer to this information as '*file meta-data*' to distinguish it from asset or failure related meta-data.

Asset failure data (particularly from actual operating plants) could be used nefariously by commercial competitors if the aggregated data (and meta-data) is overly specific. Asset owners often do not want to be identified as having equipment that fails and manufacturers do not want alternative suppliers from being able to promote their own equipment as superior. Fortunately, corporate identifiers are not relevant for developing and testing the performance of prognostic models (only the equipment/site being modeled). By removing these corporate identifiers, it is hoped that failure datasets will more readily be made available by industry.

2.3. Prognostic Data Library

A discussed in the previous section, datasets are only useful when they can be interpreted correctly. Therefore, a need was identified for a mechanism to, not only pool and share prognostics datasets with interested parties, but also ensure that relevant metadata was provided with those datasets. It is envisaged that these needs could be met using a virtual data library of suitable cataloged and appropriately classified prognostic datasets.

This prognostic data library would thus need to be able to: receive datasets and capture their associated metadata, for any type of engineering asset, in a way that obfuscated any commercial identifiers; store (and manage) the uploaded datasets and associated meta-data; be readily searchable by users in any location using commonly available technology; and provide the desired datasets and their meta-data in an easy to manage and consistent format.

2.4. Project Success Outcomes

This project will be deemed to be successful if the Prognostics Data Library (PDL) is being used by stakeholders. More specifically, the following 5 year success measures have been defined:

1. 20 research institutions or academics contributing datasets;
2. 20 companies contributing datasets;
3. 500 datasets downloaded;
4. 150 citations for contributing researchers.

Partial success will be defined by 2 or more of these measures being achieved.

3. METHODOLOGY

3.1. Stakeholder Identification

This project is being implemented using an Agile Methodology in which requirements and solutions are evolved through constant interaction between the developers and potential users of the system (Paulk, 2002). Prior to commencing any development the following stakeholder groups were identified.

a) Researchers with datasets from laboratory or industry projects who want to share their data in order to gain wider recognition of their work.

b) Researchers seeking data to test system health models they have developed.

c) Industry players with data sets who will benefit from improved prognostic models about specific equipment types and are willing to make their data available, anonymously if necessary, to accelerate the creation of these models.

In order to be successful, this project will need to ensure the needs of each distinctive group are sufficiently well accommodated. The other stakeholder is the University of Western Australia (UWA) System Health Laboratory, which is funding the project, and various UWA IT departments, who are involved in managing the hosting for the system.

3.2. Agile Methodology

The five steps making up the Agile Development Methodology are shown in Figure 1. This process is iterated continuously throughout the project.

As this process is continual and involves iteration, it is not feasible to document the outcome of each cycle. However, a number of key stages, as defined by their deliverables, can be identified:

1. Preliminary work – A basic set of requirements and design approach was identified, along with a mock-up of key user interfaces. This mockup was then used to initiate further discussion.
2. Local basic functioning prototype – A basic prototype of the system was developed and basic features, such as uploading/downloading files, could be tested by users on the UWA intranet.
3. Local more completely functioning prototype – Additional features were added to allow users to enter all required meta-data, search or browse files, and enter feedback data.

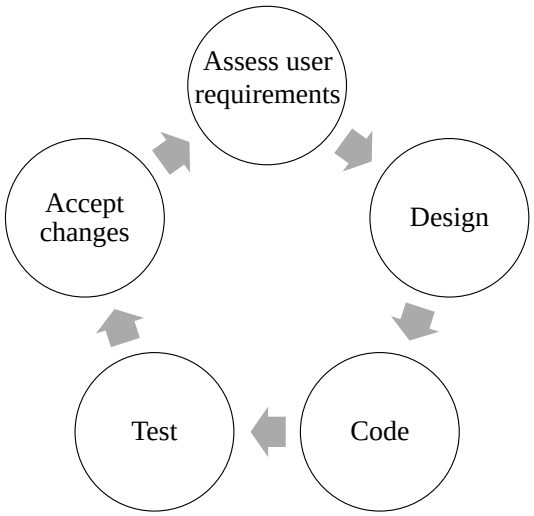


Figure 1: Steps in an Agile Development Methodology.

4. Externally accessible PDL – A web-based PDL that can be accessed from outside of the UWA domain. This initially has the same functionality as the final local prototype, along with security and user registration functions, but is made available to a limited set of external users.
5. Public PDL – Web-based PDL that is widely accessible and searchable via the internet. Functionality is enhanced based on feedback from user testing.
6. Enhanced PDL – Continually improving PDL with added functionality.

Work to date has completed stages 1-3 and stage 4 is underway. The results of the development process thus far are described in the following section.

4. RESULTS

4.1. User and Stakeholder Requirements

Key user requirement can be identified for each group of users. These are presented in Table 1.

Academics with datasets	**Researchers needing datasets**
• Generate citations for their prior work. • Benchmark their work against others using the same dataset. • Be recognized for sharing their data when applying for grants or promotions.	• Find quality datasets easily and quickly. • Spend time improving models rather than collecting data. • Datasets to be free so they could download as many sets as were suitable. • View the meta-data and summary statistics prior to downloading data. • Benchmark their work against others using the same dataset. • Datasets need to be in the same format so that recoding is not required each time. • Know what issues prior users had identified with that dataset.
Industry with datasets	
• Maintain corporate confidentiality. • Have researchers work with their data without significant additional expenditure to the business. • Know when users had developed new/improved models on their datasets.	

Table 1: Key User Requirements.

4.2. System requirements

By reviewing the needs of each user group the following high level requirements have been identified. The PDL should:

- Be useable by persons anywhere in the world without any experience of using databases or any particular programming language;
- Accommodate most (preferably all) types and formats of raw datasets;
- Accept data collected from any type of engineering asset, thus resulting in highly variable meta-data;
- Encourage users to contribute high quality data and meta-data without making the process of doing so overly onerous or time-consuming;
- Ensure that the appropriate datasets can be found when searched for, so that they are appropriately utilized;
- Track the suitability of particular datasets for prognostic modeling so that good datasets can be reused, and researchers do not waste time using poor quality datasets;
- Secure data appropriately and, at a minimum, maintain its original integrity;
- Provide instructions about how the use of the data is to be recognized or cited;

- Provide researchers with sufficient meta-data about a dataset so that it can be utilized easily and appropriately;
- Provide dataset files in a consistent format; and
- Ensure confidentiality of commercially sensitive data.

Optional features being considered include: data munging, data cleansing, collating data from multiple data-sources into a single new dataset, extrapolating sparse data or aggregating overly fine data, pre-processing data using existing tools and exporting results into a new dataset (e.g. spectral analysis, basic statistical reports), and tracking usage of datasets and users.

4.3. System architecture

The PDL is being developed using a database backend and web-delivered user-interface. Datafiles are stored as flat files, most of which are currently in csv format. It is envisaged that an additional data processing and extraction layer could be incorporated in the future to enable data to be stored in a wider range of formats (e.g. external databases, compressed binary files). As the only regular interaction required with the files is uploading and downloading, a traditional file structure is deemed sufficient. Finally, although the total amount of data is expected to be quite small (<200 datasets, 1000 registered users or 3 concurrent users), a database schema has been developed that will be able to accommodate significantly more files should the requirement arise.

4.4. Back-end Details

The PDL database stores all meta-data (asset and file) in a number of sequential, normalized databases. Examples of some of the tables are shown in Table 2.

For simplicity, the PDL backend has been implemented using an Access Database; however it may be migrated to a Microsoft SQLServer or other database server such as MySQL or PostgreSQL in the future.

4.5. User Interface

To ensure that users can interact with the PDL from any location around the world, the user interface was developed as a series of webpages. These were coded directly in a mixture of HTML, JavaScript and PHP with predefined CSS stylesheets.

User interfaces accommodate the following PDL specific tasks:

a) Uploading data and entering meta-data;

b) Data searching, preview and download;

c) Reviewing datasets and providing feedback.

Label	Description
Registered User data	Each user or contributor to the database will be registered. (User data will stored in a separate database to the file data.)
File data	Meta data describing the file such as what citation should be used, size of the data file, delimiter and other information relating to opening the file.
File contents	Details of the data in each column including a descriptor, data type, data source, processing details.
Equipment/ asset static data	Information describing the source for the data: type of equipment, type of industry, taxonomy level of prognostics prediction (e.g. system, sub-system, individual asset, component, part).
Summary statistics of data columns	Statistics for each column (e.g. statistical descriptors, number of categories etc).
Data set ratings	Stores user feedback and statistics for the number of downloads and citations.
Known identifiers	Key terms to be removed from any datasets and metadata such as company names, part names/ numbers and site names.

Table 2: Examples of tables in the PDL

The PDL website requires a significant level of interaction with the underlying data that is stored in a remote database. This database interaction must be done by the web server and therefore could potentially cause the user interface to seem slow and unresponsive. Traditionally, either all possible data had to be sourced prior to the page first loading, or pages were repeatedly reloaded when any new or additional data was required. Both options would result in the user 'waiting' for long periods of time whilst the pages retrieved and updated the web page with new information.

To overcome this, all data interaction use Ajax function calls to interact with the remote database. Ajax (short for Asynchronous JavaScript and XML) is a set of web-development techniques that can send and retrieve data in the background, whilst maintaining the display and behaviour of the existing page. When new data is available, only the relevant parts of the page are updated. In the PDL, this allows a user to scroll through a list of potential files and view the associated meta-data in real-time. When uploading metadata, the PDL is able to provide pre-populated selection boxes that continually adjust according to the user's prior selections, thus insuring highly relevant data that is easy to provide by merely selecting the best option. Freetext input is limited, minimizing the need for checking spelling and potential language vagaries.

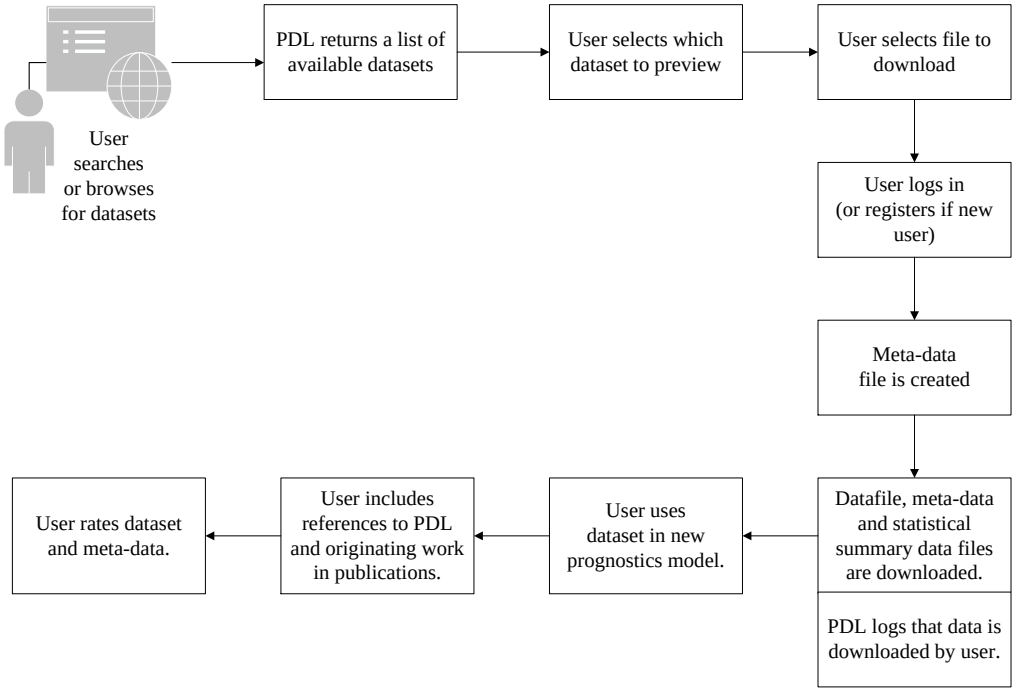


Figure 2: Process of searching for and using data in the prognostics database.

4.5.1. File Upload and Meta-Data Entry

The process of entering metadata is delivered via a web-based wizard. Only registered PDL users can submit files. The process for uploading to and using data in the PDL is shown in Figure 2.

Users are requested to enter meta-data about: the file itself and the citation to use; the rows/columns within the file; information about the asset/assets from which the data is collected; operating parameters relevant to the asset/assets.

Wherever possible, users are requested to select options to describe their data, rather than enter free text. This is advantageous for the following reasons: reduces data entry errors from mistyping, misspelling or language vagaries; simplifies meta-data storage and classification; simplifies subsequent searching for appropriate datasets; reduces the chance of identifiers being stored with the data that could identify commercially sensitive information; simplifies the process of validating the meta-data as it is entered and thereby reduces the chance of malicious code corrupting the meta-data; and ensures the required meta-data is provided.

Existing classification structures have been utilized to categorize data wherever possible. For example:

- Source industry is classified according to the ANZSIC codes as published by the Australian Bureau of Statistics. This was selected because it is readily available and covered most sectors of interest to PDL users. Should a more appropriate international classification be found, codes can be easily remapped.
- Equipment is described according to classes initially described in ISO14224 and then expanded to cover a variety of industries and asset types. An example would be the Air Transport Association (ATA) Spec 2000 standard. Where no asset classification structure is available, these are created by the authors.

It is expected that the volume of additional information supplied will vary greatly, depending on the amount of information available and the specific dataset. Data files and meta-data are reviewed (primarily to check for corporate identifiers) prior to being published on the PDL. During this time, this data is stored in a separate location so that there is no chance of access to uncleansed data via the PDL interface.

4.5.2. Data Searching and File Preview

It is envisaged that metadata will be publically available. Thus any web user is able to find appropriate datafiles by browsing through all available datasets, or by performing a free text search. Datafiles are then listed by relevance or user rating. The user can then select a datafile and review its meta-data, summary statistics and a 10% sample of the dataset. Only registered users can download files. An example of the data screen resulting from a search of the database is shown

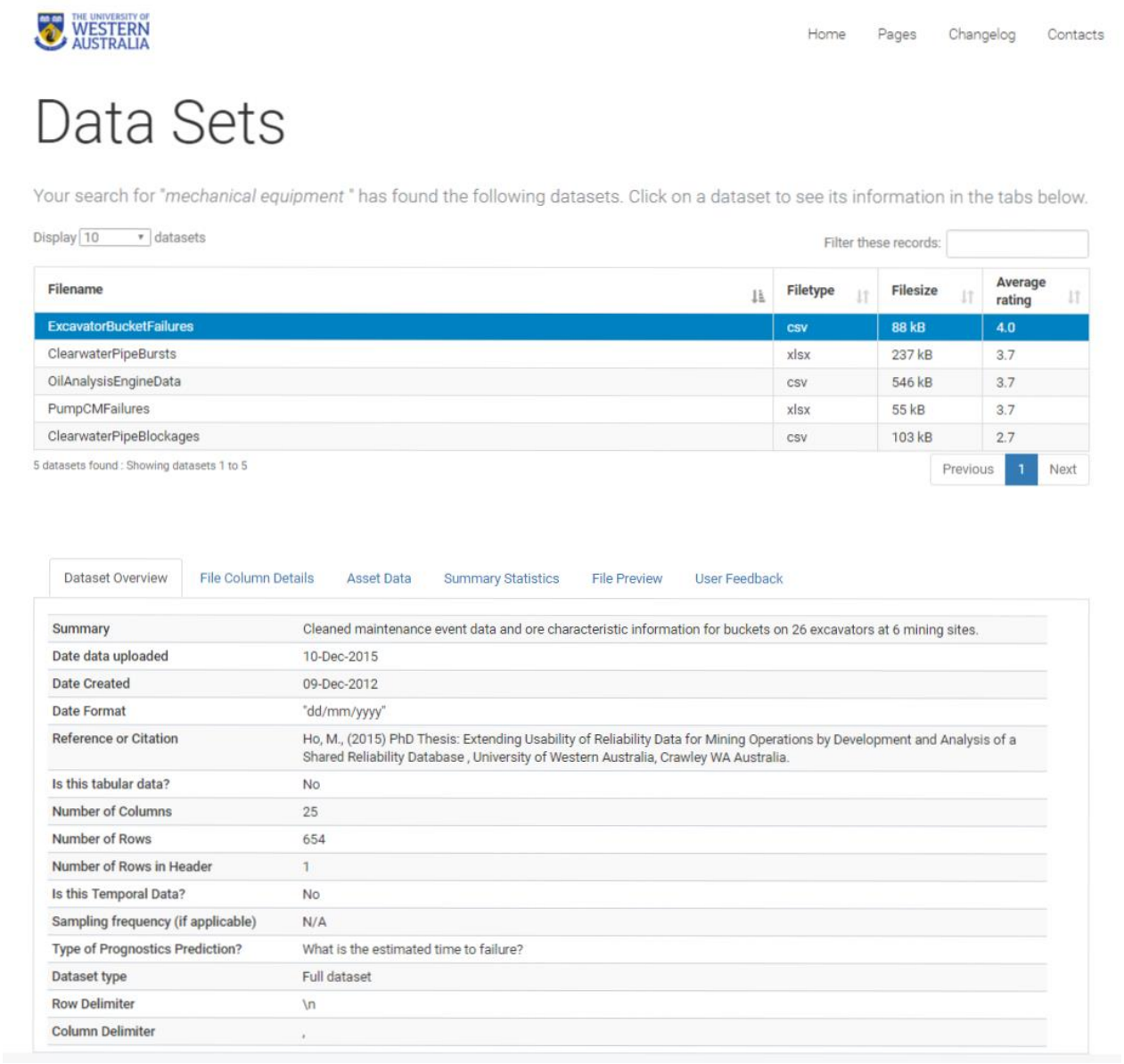


Figure 3: Example of Prognostic Data Library Data Sets page showing the result of a search in the top table and information about the data set highlights in the lower table.

in Figure 3. The top section shows the results of the search, in this case "Excavator Bucket Failures", the size and format of the file and its rating by previous users. The lower section includes five tabs of which only the contents of the first tab, Dataset Overview, is shown. The other tabs provide details of the each column in the data set, information on the asset, summary statistics of the data, the ability to preview the data, and finally a user feedback screen.

4.6. System and Data Security

We have taken a risk based approach to system and data security. The main security risks we have identified include:

a) Datasets and meta-data are intentionally used for nefarious purposes by a data owner's competitors or activists;

b) Datasets and meta-data is maliciously altered rendering it useless for prognostics;

c) Fraudulent data is supplied to the PDL.

To mitigate these risks, the following measures will be employed:

- Avoiding the storage of, and removing any remnant corporate identifiers.
- Data input validation and the minimization of free text on all data entry forms.
- 'Parking' of newly submitted information in a separate location until corporate anonymity has been verified.
- Files are stored separately to the metadata;
- Detailed citations or data collection information must be supplied with the dataset;
- Regular backups in at least 2 physically separate locations.

Although other best practice security measures will be considered, after removing any corporate identifiers the remaining information actually exists to be shared. Obfuscating corporate identifiers is the most significant risk mitigation measure used by the PDL. This is achieved by various processes.
Firstly, the meta-data entry process intentionally does **not** request the following information from users when entering meta-data:

- Asset name;
- Manufacturer;
- Model Number;
- Site.

These fields commonly include information that could be used to identify a company (e.g. MarandooCrusher2). Instead, an Asset name is automatically generated from the other meta-data that has been provided. Similarly, the original filename is replaced with a generic alternative.

Secondly, the PDL will ask users to enter corporate identifiers that must not be contained in the publically released data or meta-data. The PDL will then parse all supplied data-files and associated meta-data to ensure these keywords are not present. If they are found, keywords are replaced with randomly generated alternative classifiers (e.g. Model KY35). A list of known identifiers will also be built over time and all files checked for any entries in this list.

Thirdly, the process of removing these identifiers occurs prior to saving the data in the PDL database and file server. Finally, if users need to contact any originators of datasets, then requests will need to be made via the PDL website. The database of registered users will be kept in a different location to the metadata. The only corporate link that is retained will be a contact email address of the user uploading the data. This information will be kept in a separate database with only a numerical link retained with the meta-data.

Data owners may offer to provide access to their datasets by way of an API (Application Programming Interface). At this stage we do not envisage that this API would be published on the PDL. Instead, data queries would be created by the database as per the API, and the output of these queries would then be provided to users. The reasons for this are that by providing open access via an API: corporate anonymity could not be ensured, data would not be supplied same format, datasets would be different each time and therefore could not be rated, and finally, it would not be feasible for the PDL to publish meta-data and statistics on every possible dataset derived from the API. This approach may be reconsidered in the future.

4.7. User Feedback

Although the PDL administrators will be reviewing the datasets before they are made available on the PDL, the purposes of this review is only to: (a) remove corporate identifiers, (b) rearrange the data-file into a consistent file and (c) perform basic statistical analysis on the dataset. The administrators will not be reviewing the data files or for 'prognostic usefulness'. The best measure for prognostic quality will be user ratings. After registered users have downloaded and used the dataset, they will be requested to enter both textual feedback as well as provide three separate ratings out of five pertaining to the:

1. Quality of data;
2. Quality of meta-data;
3. Suitability for Prognostics.

They will also be asked to select what type of prognostic model was built using the dataset. Ratings will be published for registered and non-registered users to view. Other registered users with then be able to 'like' this feedback or reply to the comments. We anticipate that some useful information will come out of analysis of which datasets are used most and least often to assist our collective understanding of what makes a "good" data set for prognostics.

5. FUTURE SYSTEM ENHANCEMENTS

Future work will involve enhancements to the system that would perform tasks such as data cleansing or preprocessing. It is envisaged that the system would increasingly be able to automatically extract data from uploaded files of various types (e.g. generic binary), preprocess the data as required and then downloaded the required data selections into new files that could immediately be used for prognostic modeling or system health analysis. A webpage will be incorporated into the site where users can propose suggestions for system enhancements such as prognostic model performance metrics, enhanced searching functionality, and downloadable modules for capturing meta-data during experiments for Labview/Python/VB.

6. FUNDING

It is expected that use of the Prognostics Data Library will be free. Funding to develop and manage the system is currently available until 2020. This includes funds to implement appropriate system and data security measures. After that time, either additional funding will be sought or a suitable industry group will be approached to take over stewardship of the system.

7. CONCLUSION

This paper presents a free database for use by registered users from research institutions and industry to pool and maximize utilization of suitable datasets for testing prognostic models. Consideration has been given in the design of the database to the asset and file meta-data necessary to support prognostic modelling. A user-based rating system will provide ongoing feedback to potential dataset consumers on the usefulness and suitability of each dataset for prognostic modelling. It is expected that highly rated datasets will be used more widely, increasing the number of citations for originating sources, and thus increasing the motivation for owners of datasets to provide good quality datasets accompanied by good quality meta-data. The success of the Prognostics Data Library will depend on the database design, as presented here, on the willingness of data owners to commit data files, and modelers to make use of these files to develop and test new prognostic models.

ACKNOWLEDGEMENTS

This work was supported by BHP Billiton's Social Investment Program, which funds the BHP Billiton Fellowship for Engineering in Remote Operations.

REFERENCES

Acoustics and Vibration Database. (2013). Acoustics and Vibration Database Retrieved 8th March, 2016, from http://data-acoustics.com/

Bechhoefer, E. (2013). Condition Based Maintenance Fault Database for Testing of Diagnostic and Prognostics Algorithms Retrieved 8th March, 2016, from http://www.mfpt.org/FaultData/FaultData.htm

CALCE. (2012). CALCE Battery Group Data Retrieved 8th March, 2016, from http://www.calce.umd.edu/batteries/data.htm

Goebel, K. (2015). PCoE Datasets Retrieved 8th March, 2016, from http://ti.arc.nasa.gov/tech/dash/pcoe/prognostic-data-repository/

Heng, A., Zhang, S., Tan, A. C. C., & Mathew, J. (2009). Rotating machinery prognostics: State of the art, challenges and opportunities. *Mechanical Systems and Signal Processing, 23*(3), 724-739. doi: http://dx.doi.org/10.1016/j.ymssp.2008.06.009

Hodkiewicz, M. R., Kelly, P., Sikorska, J. Z., & Gouws, L. (2006). *A framework to assess data quality for reliability variables*. Paper presented at the World Congress on Engineering Asset Management (WCEAM), Gold Coast, Australia.

IEEE. (2012). IEEE PHM 2012 Data Challenge Retrieved 11th March 2016, 2016, from http://www.femto-st.fr/en/Research-departments/AS2M/Research-groups/PHM/IEEE-PHM-2012-Data-challenge.php

IEEE. (2014). IEEE PHM Data Challenge Retrieved 11th March 2016, 2016, from http://eng.fclab.fr/ieee-phm-2014-data-challenge/

Jardine, A. K. S., Lin, D., & Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing, 20*(7), 1483-1510. doi: http://dx.doi.org/10.1016/j.ymssp.2005.09.012

Jouin, M., Gouriveau, R., Hissel, D., Pera, M.-C., & Zerhouni, N. (2016). Particle filter-based prognostics: Review, discussion and perspectives. *Mechanical Systems and Signal Processing, 72-73*, 2-31.

Lee, Y. W., & Strong, D. M. (2004). Knowing-Why about Data processes and Data quality. *Journal of Management Information Systems, 20*(3), 13-39.

Lichman, M. (2013). UCI Machine Learning Repository Retrieved 8th March, 2016, from http://archive.ics.uci.edu/ml

MaHeMM. (2009). Railway Turnout Systems Retrieved 8th March, 2016, from http://www.aiu.edu.tr/staff/fatih.camci/datasets.html

MIT. (2016). MIT Big Data Initiative Retrieved 8th March, 2016, from http://bigdata.csail.mit.edu/

NASA. (2008). PHM08 Prognostics Data Challenge Dataset 2008. Retrieved 11th March 2016, 2016, from http://ti.arc.nasa.gov/c/13/

NASA. (2016). NASA DASHlink Retrieved 8th March, 2016, from https://c3.nasa.gov/dashlink/resources/

Paulk, M. C. (2002). Agile methodologies and process discipline. *Institute for Software Research*, 3.

PHM Society. (2009). 2009 PHM Challenge Competition Data Set 2009. Retrieved 11th March 2016 2016, from https://www.phmsociety.org/references/datasets

PHM Society. (2010). 2010 PHM Society Conference Data Challenge Retrieved 11th March 2016, 2016, from https://www.phmsociety.org/competition/phm/10

PHM Society. (2011). 2011 PHM Challenge Competition Data Set Retrieved 11th March 2016, 2016, from https://www.phmsociety.org/competition/phm/11

Sankararaman, S. (2015). Significance, interpretation, and quantification of uncertainty in prognostics and remaining useful life prediction. *Mechanical Systems and Signal Processing, 52-53*, 228-247.

Sikorska, J. Z., Hodkiewicz, M., & Ma, L. (2011). Prognostic modelling options for remaining useful life

estimation by industry. *Mechanical Systems and Signal Processing, 25*(5), 1803-1836. doi: 10.1016/j.ymssp.2010.11.018

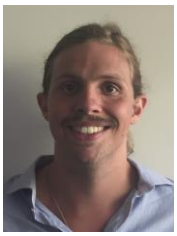
Vachtsevanos, G., Lewis, F., Roemer, M., Hess, A., & Wu, B. (2006). *Intelligent Fault Diagnosis and Prognosis for Engineering Systems*: John Wiley & Sons, Inc.

BIOGRAPHIES

Joanna Z. Sikorska received her B.E. (1st class Hons) and Ph.D. degrees in Mechanical Engineering from the University of Western Australia in 1995 and 2005, respectively. She has worked in various roles for Shell Australia and Imes Group, and has contracted to numerous organisations. Her specialisations include reliability analysis, data mining, data driven decision making and acoustic emission monitoring of rotating machinery. From 2008 until 2011 Joanna was the Technical Chair of the Asset Management Council's annual conference. Since 2007, she has run a business with her partner that designs and manufactures advanced electronics to help asset owners make better data driven decisions about their assets.

Melinda R. Hodkiewicz is the BHP Billiton Fellow for Engineering for Remote Operations at the University of Western Australia (UWA). She has a BA(Hons) in Metallurgy and Science of Materials from Oxford University in 1985, and a Ph.D. in Mechanical Engineering from the University of Western Australia in 2004. Prior to her Ph.D she worked in industry in Operations and Maintenance roles. She now leads the System Health Laboratory at UWA and works in the areas of Asset Health, Maintenance and Safety. She is a Chartered Engineer, a Member of the Institute of Materials, Minerals and Mining (IOM3) and the Asset Management Council. In 2016 she was awarded the MESA Medal for services to Asset Management.

Ashwin D'Cruz has a BPhil (Hons) in Engineering Science (specializing in Electrical and Electronic Engineering) and Computer Science from the University of Western Australia (UWA) obtained in 2016. He currently works as a research assistant with the System Health Lab at UWA in the areas of Prognostics, Electronics and Programming. His current research interests include signal processing, machine learning and music analysis.



Lachlan C. Astfalck is a PhD Candidate from The University of Western Australia working in the System Health Laboratory. He recieved is B.Eng (Hons) degree (1st class) from The University of Western Australia in 2014 and graduated as valedictorian of his class. His current research focuses on the industrial implementation of prognostic systems, with particular focus on remote systems.

Adrian Keating (M'90-SM'07) was born Melbourne, Australia in 1967. He received his B.E. (Hon) and Ph.D. (Photonics) degrees in electrical and electronic engineering from the University of Melbourne, Australia in 1990 and 1995, respectively. Since 1996 he has worked at NTT Research Labs (Musashino-shi, Japan), the University of California, Santa Barbara and at Calient Networks as the Fiber Optics Technology Manager. He joined the School of Electrical, Electronic and Computer Engineering at the University of Western Australia (UWA) in 2004 and later the School of Mechanical Engineering where he is currently an Associate Professor. His current research activities are in infrared optics sensors, sensing systems, optical microelectro-mechanical systems (MEMS), and porous silicon based sensor technologies.