

Materials Data Centre – Interim Report on User Requirements 31st July 2009

*Charlotte Rimer and Philippa Reed
Engineering Materials Research Group
School of Engineering Sciences
University of Southampton
Highfield, Southampton, SO17 1NY*

Background:

The Materials Data Centre project aims to establish an OAI-compliant Materials Data Centre designed to promote data capture and systems interoperability in the engineering materials domain. The work will extend existing DDLs in the engineering sector to incorporate standards-compliant schemas and ontologies. The work will also enhance the fractography repository at <http://www.fract.ses.soton.ac.uk>.

The work will be undertaken as a 3 year PhD project, starting 1st August 2009.

Why do we need such a facility?

The engineering sector (both research and industry) currently makes significant investments in generating materials test data. Although established testing standards specify what data needs to be recorded, this data is often not recorded, or may be incompletely recorded due to:

- (1) The data lifespan being longer than that of the storage facilities (computers, programmes etc.)
- (2) More "important" tasks taking precedence over data storage (or conservation), such as analyzing the results and writing reports or papers.
- (3) Concerns about security or IPR issues

Delays in storing the test data may also affect the quality of the information as the data become less accessible and their provenance (meaning records about origin, authenticity, and history of alterations) becomes less reliable or complete.

As surveys such as Material Information on the Internet (<http://www.brinell.kth.se/part1.html>) indicate, in the few cases where data are actually conserved, they are rarely made accessible to the wider community, often miss key information, and are not easy to aggregate with other complementary data sets or facilities. This all means we don't share data effectively, or aren't able to compare and contrast our findings in useful ways (e.g. data mining, or checking the accuracy of our own data, or finding interesting areas of new data space to explore)

The Materials Data Centre (www.materialsdatacentre.com) is one of a series of initiatives in which the University of Southampton aims to promote the capture and conservation of data in the engineering sciences. Similar data centres already exist in other domains. Such facilities allow the universal storage of data in a way that improves and simplifies access, while the

application of modern communication technologies improves data storage and data searching, shortening the processing time for both.

The primary objective of the Materials Data Centre is to allow an ever increasing body of materials data to be accumulated that is of high quality and reliable provenance, with appropriate levels of user access depending on the origin of the data.

Overview:

As part of the MDC project an assessment of the user requirements has been made through the analysis of questionnaire results (the questionnaire is now available at <http://www.materialsdatacentre.com/questionnaire.html> and we would welcome many more responses). Responses at this stage indicate that the user requirements are both wide ranging and specific to the individual. There is a predominantly positive response to the concept of sharing authenticated and enriched data, which can be linked to material condition (e.g. including micrographs). However shared concerns include confidentiality of data (for both sponsors and clients) and storage space requirements (particularly for images). There is indication that the potential value of the MDC will be linked to how user friendly it is. The creation of this report by no means represents the end to the requirements gathering stage of this project, but is simply a starting point designed to ensure the project begins along the correct route. The initial survey aims to establish foremost the activities in which the user is engaged, then the ways in which they use collected data, and finally the individual's data management needs.

Methodology:

The primary aim of this preliminary 8 weeks internship project is to establish the user requirements of a materials data centre. To that end several methods of questioning users have been developed to achieve this aim. For each method a common set of questions was created using an iterative process. The primary stage in establishing user requirements was to establish the potential user groups themselves. This allowed questions to be created from the perspective of a user.

The questions created, finally totalling twelve, each consisted of two parts; a leading introductory part with a yes/no answer, and then a follow up part which allows the user to impart further explanation or comment. The topics for questioning flow throughout the questionnaire, and can be considered to be in three parts; data, standardisation and storage.

There were three methods of distributing the questionnaire; paper copies filled out by individuals but in a group environment, one on one interviews with academics and now a web-based questionnaire available from:

<http://www.materialsdatacentre.com/questionnaire.html>.

The web-based questionnaire will be kept open during the MDC development phase, and volunteers (recruited via the questionnaire) will be invited to trial the MDC at appropriate points.

Sample:

The sample taken is as wide ranging as possible and varied from undergraduate users to academics and some members of industry. Contacts used ranged from within the Southampton University materials research group to other universities, and contacts met at conferences. To date there are 28 user responses. As it stands currently the sample group consists of the following; eleven PhD students, 9 academics, 1 person from industry, 3 undergraduates and 4 postdoctoral researchers. The responses from the web-based questionnaire have not been incorporated into this interim report, but this ongoing user-requirements assessment will be updated on a monthly basis.

The questions:

1. Does your research focus on specific materials? Yes/No

64% of users answered yes.

Which ones and with what objectives?

Most common alloys named here were Aluminum (20%), Titanium (20%), Nickel (15%), Steel (20%) and CFRP (25%).

2. Do you already use standardised testing techniques? Yes/No

86% of users answered yes

If so which ones and would you want to specify new testing methodologies?

BSI and ASTM were the most common sources for standardisation. These predominantly included fatigue, tensile, hardness, surface roughness, three point bending, four point bending, long and short crack tests and Charpy impact.

3. Does your research depend on working with experimental data? Yes/No

75% of users answered yes.

Measured (observational, experimental) or derived and for what purpose e.g. FEA, modeling, data mining?

The most common answer by far here, with 54% of users citing it as a purpose, was FEA. Most users were interested in measured data, but there were also a proportion (18%) interested in derived data.

4. Do you ever use data from other sources? Yes/No

71% of users answered yes

What is your experience of the availability, quality, and usefulness of this sort of data?

Essentially the outcome of this question showed that the availability, quality and usefulness of this data was highly variable. It is fair to say that for the most part that simple data (common values) was much easier to find than more difficult data (see appendix of comments), such as high quality images.

5. Is there ever a requirement to augment your data from other sources, such as literature and web-enabled databases? Yes/No

61% of users answered yes

What are the circumstances?

The most common answers here were that data augmentation is used either in the literature review, or as a comparison for results achieved by the user.

6. Would you find it useful to have a universal method of storing data? Yes/No

71% of users answered yes

What type of data would you find it useful to store and to whom would you like to make it available?

It is fair to say users were unsure of the possibilities and viability here. To get a clear list of deliverables a much larger sample size is required, as there is currently little commonality of requirements established.

7. Have you ever needed to use experimental data from earlier years and/or previous projects? Yes/No

75% of users answered yes

What are your experiences locating and using such data?

Overwhelming response (57%) reporting finding this data was difficult and time consuming.

8. Have you ever shared your data either on request or in the context of collaboration? Yes/No

61% of users answered yes

What have been your experiences and did/would you have any reservations about sharing your data?

The major concerns here were confidentiality and acknowledgement, although there were also concerns over the commercial value of data. Mutually beneficial relationship required.

9. Do you have security concerns for the data you would be adding to the data centre? Yes/No

46% of users answered yes

What are the major issues?

Commercial sensitivity was a major concern in this area, however sharing published data was not considered an issue.

10. Do you consider conservation of experimental data worthwhile? Yes/No

89% of users answered yes

What do you think would be the advantages and disadvantages?

Here users wanted the facility in order to back-up their work, but also believed it would prevent the repetition of work and therefore create greater continuity. Concerns were related to the cost and space requirements.

11. Would you be prepared to trial the data centre? Yes/No

75% of users answered yes

If you supplied data to the centre how do you envisage it being used and what constraints would you need to apply?

Most users agreed there would need to be some sort of restriction on data access, however to get a clearer view of the uses envisaged again a greater sample size is required.

12. In general would you like to see the engineering materials sector moving towards a more consistent/standardised data layout? Yes/No

86% of users answered yes

If this is the case what would you ideally like to find in a data centre and what are your major concerns as a user?

Once again for this question there were lots of concerns over feasibility, with the majority of users agreeing it's a good idea in an ideal world. There were also concerns that standardisation would inhibit progression. (Fuller listings can be found in the appendix).

Appendix: Full listing of question responses.

Specific materials:

- (phd) WC-Co
- (phd) Al-alloys
- (phd) Hydroxyapatite titanium alloy
- (undergraduate) Acrylic bone cement
- (phd) 12 Cr ferric heat resistant steels
- (PostDoc) Ni-superalloys
- (phd) CFRP
- (phd) Carbon Fibre Composites
- (Academic) Ti alloy, Ni superalloy, Al, Carbon Fibre, pressure vessel steels
- (academic) Al, steel alloys, single crystals DS (creep fatigue, fracture)
- (academic) Mostly metallics in engine applications (Ni based – aeroengines, steels – power generation, Al-Si – pistons)
- (academic) Al alloy – microstructural properties, plastic deformation, expected failure mechanisms
- (phd) Bone/lung tissue
- (academic) hard coatings, shape-alloys (based on properties/cost)
- (PostDoc) dyes, silicon solar cells
- (PostDoc) Ti and CoCr alloys
- (Phd) Inert bioceramic
- (undergraduate) ceramic with CFR-PEEK
- (industry) aluminium, steel, CFRP.
- (phd) Ti, CoCr, UHMWPE, PMMA.

Standardised tests:

- (phd) British standards: fatigue, tensile, hardness, surface roughness.
- (phd) High temp testing methodologies
- (phd) British standards; implant testing
- 3 point bending, force spectroscopy, nano indentation
- (PostDoc) BS and ASTM: long-crack tests, Bespoke (comparable in-house): short-crack tests
- (academic) Many ASTM and BSi mechanical testing methods
- (phd) Implementing new standards in a field governed by commercialism is difficult
- (Academic) UKAS accredited x-ray stress measurement unit. Developing composites test and evaluation facility, certifying carbon fibre properties. New methods of certifying composite test pieces and structures required.
- (academic) Astm, BSI
- (academic) mostly fatigue and tensile and toughness and hardness, variants on testing approaches and develop methodologies.
- (academic) would use standardised testing if given standard samples – adapting to situation.

- (academic) tensile, static, 3 point, 4 point, flexural, fatigue, tensile tension-tension.
- (academic) ASTM
- (phd) 3 point bending, force spectroscopy, nano indentation
- (academic) standard tensile test, Charpy impact, composition, wear test have no European standard but are common.
- (PostDoc) Many spectra (absorption, fluorescence, time fluorescence, ellipsometry, transmittance)
- (Phd) 4 point bend, light/sem microscopy → FEA
- (undergraduate) British Standard materials testing
- (industry) ASTM, British Standards, coefficient of linear thermal expansion (Non standard)
- (phd) 4 point bend, tensile test, compression test.

Data:

- (PostDoc) Measured for data analysis
- (phd) Experimental data image analysis with the potential for FEA
- (phd) FE-measured (experimental) and derived from observations
- (academic) Wide ranging purposes including 3D visualisation, FEA, modelling and data mining
- (phd) Measured/derived crack length
- (undergraduate) Measured (TEM, SEM...)
- Measured data from experiments
- (phd) Measured data from experiments and FEA
- (phd) Measure and compare with modelling results
- (phd) Using Matlab, ANSYS, maple and solidworks
- (academic) FEA and modelling
- (PostDoc) Measured and derived
- (academic) Measured and derived for FEA, modelling and data mining.
- (academic) experimental and derived for FEA and modelling.
- (academic) Measured experimental data for FEA models.
- (academic) Measured and derived for data mining.
- (academic) Mechanical properties/ testing and hardness properties for image data e.g. SEM
- (phd) measured data from experiments
- (academic) measured – wouldn't trust models. Would be useful for FEA, massive lack of data for this and FEA limited by data input.
- (PostDoc) Experimental (e.g. spectroscopic measurements, IV measurements) and modelling (ellipsometry, time fluorescence decays)
- (PostDoc) Muscle forces derived from patient data to drive FEA models.
- (PostDoc) Measured for FEA
- (phd) Measured – statistically derived – FEA

- (undergraduate) experimental
- (phd) Gait analysis data for the purpose of FEA.
- (industry) Experimentally measured, FEA used to interpret results.
- (academic) All types especially input into FEA.
- (phd) FEA and other computational techniques.

Other data sources:

- (phd) Quality dependent on information given to provider
- (phd) From a specific proposal quality is excellent
- (academic) Quality highly variable
- (PostDoc) Not easily available, some proprietary issues, quality variable, usefulness depends on quality
- Databases are available for academic papers but not well categorised
- (phd) Published papers used
- (phd) Older sources difficult to find and often incomplete
- (academic) Rolls Royce COMMIT database is high integrity.
- (phd) Very difficult to locate data in the literature that is relevant to material type and type of test under consideration
- (academic) Invalidated and usually not trustworthy.
- (academic) Variable – some tracing issues. Need quality assurance, the more background data on the material the better. All the information required is not always present in the paper.
- (academic) there are too many different alloy possibilities so it's better to repeat tests.
- (academic) Good experience because needs are simple – numbers can be approximate because of the sensitivity analysis.
- (academic) Standard materials properties – readily available and self contained.
- (academic) try to get whatever is available on the internet, published papers etc – especially true for fatigue data under specific conditions. Would be useful to have lost data back. The data exists but can't be found. Poor funding for publishing a paper about data. Data is in commercial hands.
- (PostDoc) Spectra from dye databases available in html format. Optical constants available from commercial software.
- (PostDoc) Good availability, quality and usefulness.
- (phd) quality is unknown or from peer review journals.
- (industry) not much is available, the quality is questionable and usefulness can be low due to the large ranges quoted.
- (academic) things which can't be measured easily e.g. coefficient of thermal expansion.
- (phd) Tediously trawl literature for useful figures.

Augmentation:

- (phd) Determine baseline characteristics and compare expected results.
- To improve sample sizes
- (PostDoc) Compare and contrast results
- (academic) Research always involves some cross-checking
- (PostDoc) For a literature review
- (academic) From literature for FE modelling
- (academic) Augment when there is unavailability of good data.
- (academic) Uses literature, might need to interpolate data or make assumptions on material without having to do the test.
- (academic) Only as comparison/discussion.
- (academic) potentially. If data was readily available would be more inclined to use it. Ease and availability.
- (academic) Not a strong requirement currently but can see that changing.
- (phd) database would be useful in providing large sample sizes
- (academic) especially on consultancy activities and if you want to be more quantitative – growing trend towards this due to guarantee period.
- (PostDoc) mainly dye databases with spectra or solar cell efficiencies, or optical constants.
- (PostDoc) Gathering data (loading, implant positioning, material properties, friction). For FEA inputting strengths of materials, failure mechanisms.
- (industry) to sanity check experimental values.

Universal method:

- (PostDoc) Not sure
- (academic) The possibilities are complex to understand. Data handling is an important topic. Can't personally see a good solution.
- Depends on data
- (phd) Storing raw data from CT scans (4 TB), available with in group and external applicants
- (academic) Property measurements, images, model formulations, model outputs
- (PostDoc) Would depend how much extra information would have to be recorded to make it useful to others. Post processed information. Access controlled by request or perhaps output e.g. graphs rather than full data access (cost of production, could charge for access)
- Make data available to in house researchers
- Sample information, raw data and processed experimental data
- (phd) Shared set-up

- (phd) Processed data with relationships. Also share data between individuals working on the same or similar projects.
- (academic) mechanical property and thermal data. There is a composite database in the US.
- (academic) yes-validated data with known pedigree.
- (academic) hesitant. "universal" may be too general to be useful – requirement for tailoring!
- (academic) yes. Good to have data to compare but won't be exact as there are too many different alloys.
- (academic) Basic mechanical property data. Dynamic and static.
- (phd) sample information, raw and processed experimental data.
- (academic) don't want to store data only for the software to change. Any sort of data related to service performance (properties) or fatigue and wear data. The more data the better, helps designers to design things that don't fail – reduce safety factor, save money.
- (PostDoc) Spectropic data, IV curves – make them available to academic staff and other researchers.
- (postDoc) Good for results and inputs for FE but not for patient data.
- (undergraduate) material strength data, probabilistic data.
- (phd) gait analysis data for users
- (industry) any measured data. Available to the UoS research community. If made externally available would need to charge and ensure data quality.
- (phd) As a repository of computed tomography data would be useful. Also catalogued store of previous model results.

Past projects:

- (phd) Difficult to find the correct, useful data.
- Would be easier if deposited in a database together with accompanying information-sample description and test parameters
- (undergraduate) Time consuming, depends on quality, difficult if there are contradictions within teams of researchers.
- (PostDoc) Data, not too bad. Images, poor, photographic images often degraded or lost.
- (academic) Variable if outside own research
- (phd) Very difficult if author is not known personally
- (phd) Good if data format is the same as currently used
- (PostDoc) Make sure data is transferable within one person's research area, but difficult to access other's data
- (academic) Haphazard experiences in this area.
- (academic) difficult.
- (academic) Variable. Can be hard- everyone uses different filing systems, data formats change.

- (academic) Problems locating it.
- (academic) data easily available in lab books.
- (phd) would be much easier if deposited in a database together with accompanying information- sample description and test parameters.
- (academic) struggle – not searchable, sometimes destroyed.
- (PostDoc) Accessing own data is a lengthy process, accessing other people's data is impossible.
- (PostDoc) Located using literature search, not always reliable.
- (phd) difficult sometimes
- (industry) usually in previous log book and papers but finding it is time consuming.
- (academic) near impossible, especially with raw data.
- (phd) Lucky if you receive information.

Collaboration:

- (PostDoc) Fine if properly acknowledged
- (phd) Size is the main issue (typical scan 7GB)
- (phd) Only processed images
- (academic) Depends on novelty, commercial value and competitiveness
- (PostDoc) Sharing data on precondition of receiving material
- Good experiences, no reservations on sharing of published data
- (phd) There is potential to share between a relevant company and the university
- (academic) share with industrial collaborators.
- (academic) difficulty in different formats.
- (academic) worry that data is "expensive" or that industry won't give credit appropriately.
- (academic) with other universities no reservations in collaboration since each party tends to only be interested in their section of the project.
- (academic) No real reservations. Important issue is how the data is used.
- (academic) no reservations. Straight forward data – easy to transfer
- (phd) good experiences, no reservations on sharing if published data.
- (academic) if on a project together. Share as long as it doesn't conflict with confidentiality.
- (PostDoc) if published and acknowledged no problems.
- (PostDoc) No reservations – only shared with established contacts/collaborators or with people on the same project.
- (PostDoc) Get useful feedback from others, no reservations(phd) working with industrial partners causes a problem, require data protection.

Security concerns:

- (phd) Yes due to external funding
- No if published
- (PostDoc) Possible proprietary issues from sponsor. Cost of production and time. Others could publish data before the author
- (academic) Commercial sensitivity of sponsors in particular
- (phd) Only available to external parties through application and direct permission
- (phd) Plagiarism
- (PostDoc) Risk of inappropriate use of data without author's knowledge. Also some data has potential commercial sensitivity.
- (academic) liability.
- (academic) Funder issues
- (academic) Not for published data.
- (academic) Issues with sponsors, ownership of data.
- (academic) Other's would as it is other people's data.
- (phd) not if published
- (academic) confidentiality (ok if informed), security (changing values without authority), ok if everyone gains
- (PostDoc) Patient confidentiality, external back ups.
- (phd) commercial confidentiality.

Conservation:

- (PostDoc) Useful only to a certain extent. Advantage of shared research activities, discussions, collaboration etc. Disadvantage of breach of data security etc.
- (phd) Advantage – back up and security. Disadvantage – lack of speed depending on access type.
- (phd) May prevent unnecessary repetition of work. Space/cost. Easy access to previous/relevant work.
- (PostDoc) Advantage – easier access (especially micrographs), other data may be available. Disadvantage – security, pre-publishing
- (undergraduate) Can see others finding. Avoids the loss of data.
- Repository of interest to community, especially researchers waiting to use experimental data for modelling or for reproducing experiments.
- (phd) Advantage – check data any time. Disadvantage – accessibility, storage space.
- (phd) It will save time in getting useful data. Researchers can do fewer experiments individually.
- (phd) Continuity of previous work. Creating additional information rather than dealing with the necessity of possibly repeating already known information.
- (phd) Follow on projects can then easily find a start point to define the work that needs to be done

- (academic) Yes for data that is expensive to get. No for simple data (personal experience). Long-term data is a big issue so yes. May be cheaper to re-do test than conserve data.
- (academic) Disadvantage – pollutes database if data is not sufficiently accurate or well characterised.
- (academic) good for future use.
- (academic) Avoids re-inventing the wheel. Checks and balances. Reveal wider ranging relationships.
- (academic) Advantage for similar alloy comparison, reference only, could create relationships that weren't exactly truthful, caution required.
- (academic) very useful as past data is often lost.
- (phd) repository of interest to community, especially researchers wanting to use experimental data for modelling or reproducing experiments.
- (academic) All experimental data is useful – shouldn't be buried in thesis.
- (PostDoc) advantage - easy access. Disadvantage – large disk space and resources.
- (PostDoc) Future generation can use it. Avoids duplicating work. In the spirit of "research community".
- (PostDoc) Advantages: access for people interested in using the data, comparison purposes. Disadvantage – storage.
- (phd) Linking postgrads research, not just discrete projects.
- (undergraduate) Advantage of future reference.
- (phd) it can be compared to new studies but it might have large space requirements. People may have varying techniques.
- (industry) for future use.
- (phd) advantage-no repetition of experiments, disadvantage-only useful if properly catalogued.

Trial:

- (phd) Used in follow on projects (if sponsor is constant), outside this sponsors permission required.
- Storage facility for extra work
- No reservations for published data
- (phd) Access for measurement of progressive damage growth from direct observation. Copy right concerns and acknowledgements.
- (academic) Consider against resource levels – CT data!
- (phd) Only as a backup facility. Access restricted to selectable users
- (PostDoc) Interested parties have discussion with author first
- (academic) depends on agreement.

- (academic) might need to restrict access of some users. Should be used in every way possible.
- (academic) probably, only use as a reference, general property. No restraints required.
- (academic) supply of data is difficult as it is not usually produced as a formalised data sheet in a thesis. Data provided must be a trustworthy, common form.
- (academic) Need to do something with data generated, data is diverse, would be useful in-group
- (phd) at first only published data, no reservations
- (academic) data should be used in design context, provides more information from which to select alloys, avoids repeating experiment.
- (PostDoc) Used for demonstrations, and not shared without permission.
- (industry) for anyone who would find it useful, can't envisage any constraints.

Standardised layout:

- (PostDoc) Lots of potential cautions
- (phd) Ideal world
- (academic) Provenance
- (phd) Access to relevant data/easy to find specific data among vast quantities.
- (PostDoc) Utopia = full data for all materials. Optimum = mandate fields for different types of data rather than all fields
- (undergraduate) Simple to use. Results dated, short loading time, history of previous searches.
- Description of sample, experiment, date. Raw experimental data and processed data.
- (phd) Test set-up (equipment used), standard used, outputs
- (phd) Concerned with data reliability.
- (phd) Concerned with data management. Key word specific search. Security, author's permission
- (phd) Processed data for comparison. Graphical data and data used to generate it. Micro structural and fracture surface images.
- (academic) Major concern is that the data isn't sufficiently well characterised or reliable e.g. heat treatment condition of alloy etc.
- (academic) standard format and ease of access.
- (academic) not sure this allows for new knowledge. Searchable in a flexible way. Provenance.
- (academic) Yes for traditional alloys only because the nature of the sector is developing and fast moving.

- (academic) Timescale is massive. Unpopulated resources are frustrating.
- (academic) Not convinced it's feasible. Researcher and designer use data differently.
- (academic) easy comparison, lots of applications aren't standardised. Ideal world – any data anyone has done – find it easily – great help.
- (phd) description of sample, experiment, date. Raw experimental data, and processed data.
- (PostDoc) major concerns will be IP publications
- (PostDoc) easy to use layout i.e. advanced search functions. Either actual data or links to where data can be found.
- (PostDoc) major concern is reliability.
- (Phd) enjoyed the “logs, blogs and pods talk”. Most work is commercially confidential.
- (phd) material properties, standard gait analysis tests from a range of people, CT scans.
- (industry) useful material, the validity of it.
- (academic) standardisation is a barrier to progression.
- (phd) ease of access uploading and downloading information.