A Material Assay Database for the Low-Background Physics Community

J.C. Loach*,†, J. Cooley**, G.A. Cox‡, A.W.P. Poon†, K. Adler**, M. Bruemmer**, K.D. Nguyen† and B. Wise**

*Shanghai Jiaotong University, Shanghai 200240, China
†Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
**Southern Methodist University, Dallas, TX 75275, USA
‡Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany

Abstract. The low-background physics community has been performing assays of material radiopurity for decades and many thousands of such measurements exist. Some are available in publications and others in databases, but most are still communicated informally. No standard data format exists for encoding material assays and there is no central location to store them. The aim of the work reported here is to address this long-standing problem, through the creation of a concise and flexible material assay data format and powerful database software to use it. A public installation of this software may serve as a long-term repository for the community's material assay data. It is available at `http://www.radiopurity.org`

Keywords: assay, dark matter, database, double beta-decay, mass spectroscopy, radioactivity
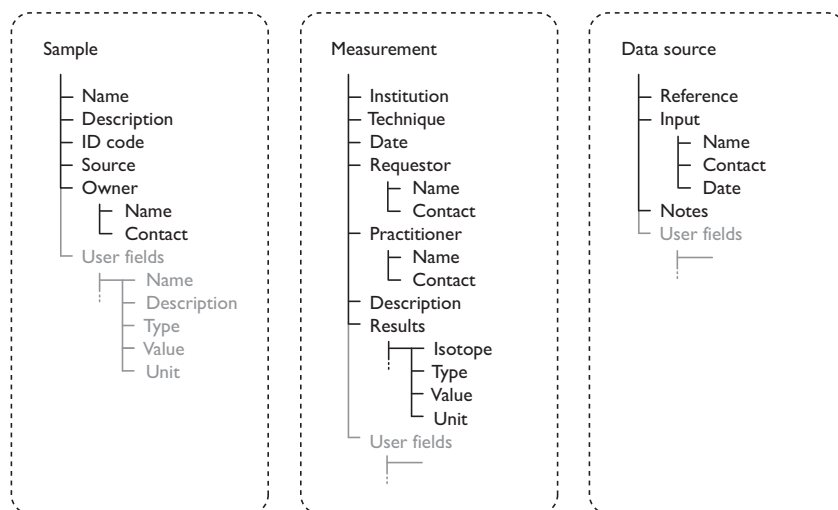
INTRODUCTION

Crucial to the performance of experiments that search for rare weak interactions, such as $0\nu\beta\beta$ decays or the interactions of WIMP dark matter, is the suppression of events due to radioactive backgrounds. This can be achieved through experimental design and though analysis techniques, but the most basic strategy is to reduce the source terms, by shielding the experiment from environmental radioactivity and, of particular relevance here, by constructing the experiment from low-radioactivity materials.

The selection of low-radioactivity materials is usually informed by existing measurements and by experimental experience. Selection is typically followed by the screening of batches of the candidate materials using techniques such as gamma counting or mass spectroscopy. These assays are precision measurements that require significant time, cost and effort, and most experiments perform large numbers of them, typically screening all of their construction materials. Judicious selection of materials, of supplier and production process as well as chemical composition, is thus an important aspect of building a low-background experiment. The efficiency of this process relies heavily upon the quality of information available to experimentalists and in particular on their access to previous measurements of similar materials. Such information increases the probability that selected materials will meet the experiment's requirements and can thus reduce the resources that are required by the screening process.

A central repository of the results of material assays is strongly motivated by arguments such as this and the authors are not the first to propose one or, even, to build one. Previous efforts such as the public database of radiopurity measurements created by the ILIAS collaboration [1] have been invaluable to the low-background physics community. But despite their usefulness, none have fully met the community's needs, principally due to the restricted size of their data sets and how awkward they are to use, maintain or augment. The work described here is the authors' attempt to address the shortcomings of the previous efforts and to provide a long-term solution.

A database system has two main components: the data format - how the data is encoded - and the implementation - the mechanics for storing and accessing the encoded data. The components are necessarily intertwined but some effort has been made in these proceedings to treat them separately and for a specific reason: because we are advocating the adoption of our data format as a standard for material assay data *independent of our particular implementation*. The standard format is the most important aspect of this work. This said, we also believe that our software, described following the data format, is more than adequate to meet the needs of the community.

Sample
├─ Name
├─ Description
├─ ID code
├─ Source
├─ Owner
│   ├─ Name
│   └─ Contact
└─ User fields
    ├─ Name
    ├─ Description
    ├─ Type
    ├─ Value
    └─ Unit

Measurement
├─ Institution
├─ Technique
├─ Date
├─ Requestor
│   ├─ Name
│   └─ Contact
├─ Practitioner
│   ├─ Name
│   └─ Contact
├─ Description
├─ Results
│   ├─ Isotope
│   ├─ Type
│   ├─ Value
│   └─ Unit
└─ User fields

Data source
├─ Reference
├─ Input
│   ├─ Name
│   ├─ Contact
│   └─ Date
├─ Notes
└─ User fields

FIGURE 1. The three top-level components of the material assay data format. Each component is represented by a JSON object.

THE MATERIAL ASSAY DATA FORMAT

The Material Assay Data Format (MADF) is intended as a standard data format for codifying assays of material radiopurity. This section is an introduction to MADF version 2.0. Unfortunately the confines of this proceedings do not allow a formal specification or even a complete description.

The aim of MADF is to capture the essence of a material assay, including information on the sample and the measurements performed on the sample. The guiding principles behind its design are as follows:

Simple The data format is flat, consisting of a single record for each assay and there are no relationships between records. This has the benefit of making the data highly portable in that any set of measurements can be trivially output to independent text files. The format has a purposely restricted number of fields and is intended to be closer in form to a structured experimental report than a line in a spreadsheet. This simplicity helps in the process of codifying data: it helps if the source information is already expressed in a structured digital format, making the process more amenable to automation because most conversions will be from a more complex format to a simpler one; and it also helps if the source information is unstructured, because it minimizes the number of decision that need to be made during data entry. Simplicity in a data format tends to make manipulations of the data more difficult but this issue is greatly eased by modern techniques for data retrieval and synthesis.

Inclusive The data format is weakly validated, in that there are minimal constraints on what can be entered in particular fields. Examples are the absence of constraints on units and the absence of requirements that measured values have errors and confidence levels have limits. In both these examples there would have been clear benefits to standardization but the data format was primarily designed to be as accommodating as possible to existing forms of data, both formal and informal; the premise is that more information is usually better and that users of the database will be sufficiently qualified to judge what is useful for their purpose. A measured value without an error is not necessarily meaningless and no one should be deterred from contributing their data by the effort of converting units. Key to this working in practice is that the user should be able to understand where the data originated and any assumptions or simplifications made during data entry. The lack of such information has been a principle shortcoming of previous systems and has been addressed in MADF with fields for contact details of individuals involved in the measurement and data entry, and for notes on the data entry.

Flexibile The flexibility of the data structure is closely related to the two items above, in that it allows the core data structure to be simple and yet makes it accommodating to those users who have need of more structure, for example for interoperability with other digital systems. The data format allows arbitrary extension though user-defined fields. A typical use of this feature would be to allow parameters to be stored which are of specific technical use to the generators of the data but not of general interest to others in the field. Examples might be the name of data files storing HPGe spectra or calibration data for an instrument.

```
"practitioner": {
  "name": "Alice",
  "contact": "alice@university.edu"
},

"results": [
    {
      "isotope": "U-238",
      "type": "measurement",
      "value": [400, 20],
      "unit": "ppb"
    },
    {
      "isotope": "Th-232",
      "type": "limit",
      "value": [100, 90],
      "unit": "ppt"
    }
],

"user": [
    {
      "name": "Data file",
      "type": "string",
      "value": "2010-A-0092.spe"
    }
]
```

FIGURE 2.   A portion of the measurement component in the material assay data format.

The data format is expressed in JavaScript Object Notation (JSON), a text-based open standard for data interchange that is widely used by internet applications. The format is language-independent and parsers are available for all major programming languages.

The MADF data structure is divided into three top-level components, each a JSON object. These store information about the sample, the measurement and the input of the data. Additional top-level fields store information such as the MADF version and the grouping (for example, the experimental collaboration) to which the measurement belongs. Figure 1 shows the structure of these components schematically.

An example of part of the measurement component is shown in Figure 2. This illustrates some of data structure's typical features:

- The `practitioner` object stores the name of the person who made the measurement and their contact details.
- The `results` array holds a list of objects representing results for single isotopes. The array can vary in length and the `value` array for each entry is interpreted according to the `type`. For U-238 in this example the `type` is 'measurement' so the first number in the `value` array is interpreted as the central value and the second as the magnitude of the symmetric limit; the `value` array could have been shortened to indicate that there was no quoted error or extended to allow for an asymmetric error. A similar interpretation applies to the entry for Th-232, where the `value` array holds a limit and confidence level.
- A single `user` field stores the name of HPGe detector spectrum data file. The file name happens to be a string but other user field types are supported.

At the time of writing almost 1000 material assays have been encoded in MADF.

THE COMMUNITY MATERIAL ASSAY DATABASE

This purpose of this section is twofold, to outline a software architecture in which MADF can be used effectively and to give a concrete example - a database system built as a central repository for community material assay data.

The software architecture is shown in Figure 3. The database engine is Apache CouchDB [2], an open-source NoSQL database that stores loosely structured data in JSON documents and uses HTTP as its primary API. CouchDB is designed for use by web applications but its HTTP API means that it is straightforward to interface with from most traditional programming languages. The database's schema-less design provides the flexibility required by MADF and
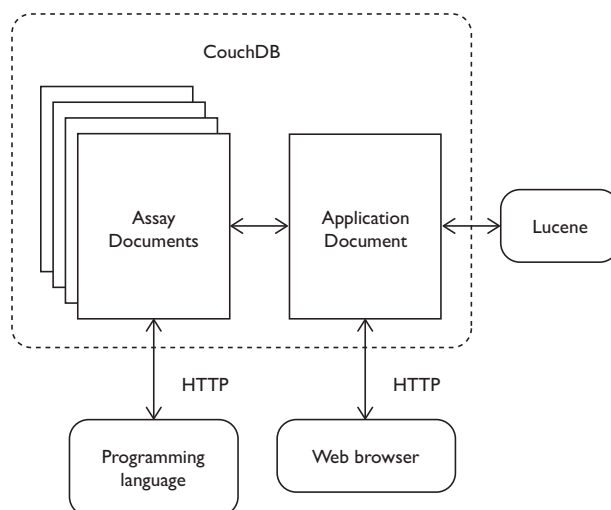
FIGURE 3. The architecture of the database application.

also the ability to store client-side web application inside JSON documents. This latter feature means that CouchDB can act both as a data store and a web server to provide a database interface to users. Plugins to CouchDB allow web applications to implement sophisticated search capabilities.

This architecture has been used to build an open-source software package to store material assay data called Persephone. This software consists of a client-side web application that respects the MADF specification and is served by CouchDB. The application allows powerful search of the information contained within the database and flexible display of the output. It also provides an input form to facilitate data entry. The application is designed to be run in a number of use cases: locally by private individuals, by institutions who perform material assays, by experimental collaborations and also as a central public database of community data with regional mirrors. It has been designed to allow straightforward transfer of data between these instances and therefore all public data to be effectively held in common. The community installation of the software can be found at `http://www.radiopurity.org` At the time of writing it contains data ported from the ILIAS database and from a number of peer-reviewed publications [3, 4, 5].

The MADF specification and the application architecture described in this section provide a long term basis for solving the problem of efficient community storage of material assay data. MADF is freely adaptable to user needs and to future applications, and the architecture provides a straightforward and highly flexible framework for developing applications that are compatible with MADF.

REFERENCES

1. ILIAS database on radiopurity of materials, `http://radiopurity.in2p3.fr/` (accessed 2013-06-14).
2. Apache CouchDB, `http://couchdb.apache.org/` (accessed 2013-06-14).
3. D.S. Leonard et al., *Nucl. Instrum. Meth. A* **591**, 490 – 509 (2008).
4. E. Aprile et al., *Astropart. Phys.* **35**, 573 – 590 (2012).
5. G. Alimonti et al., *Astropart. Phys.* **8**, 141 – 157 (1998).