Metadata Management in Scientific Computing

Eric L. Seidel[1], Gabrielle Allen[2]

[1]Department of Computer Science, City College of New York
[2]Center for Computation & Technology, Louisiana State University

Abstract

New methodologies are needed to support data sharing in virtual organizations that develop and deploy complex simulation codes on large scale HPC resources. A sophisticated categorization environment is needed that will allow the community to store, search and enhance metadata for scientific codes and the datasets they generate in an open and dynamic manner. Currently, data are often presented in a *read-only* format, distilled and curated by a select group of researchers. We envision a more open and dynamic system, where authors can publish their data in a *writeable* format, allowing users to annotate the datasets with their own comments and data. This would enable the scientific community to collaborate on a higher level than before, where researchers could for example annotate a published dataset with their citations.

In this poster, we present an alternative method of publishing codes and datasets, based on Fluidinfo[a], which is an openly writeable and social metadata engine.

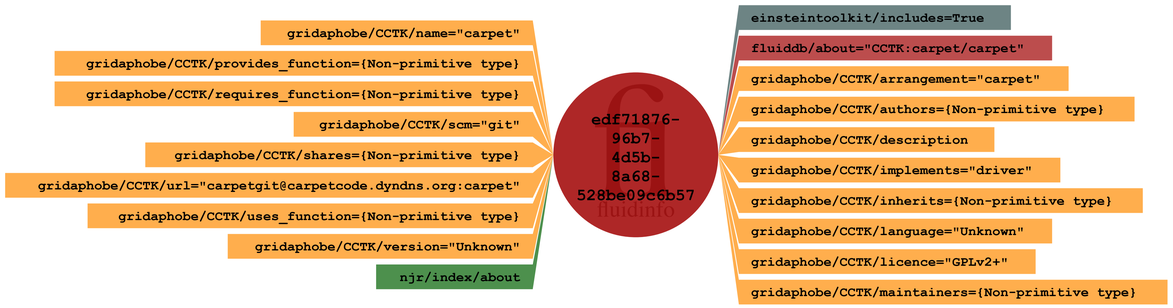

Figure: Visual representation of the Fluidinfo object for the Carpet module in the Einstein Toolkit [3]. Note the collection of tags from four different users.

Fluidinfo

Fluidinfo [4] is a schema-less datastore that encourages collaborative annotation of shared objects. It operates based on the metaphor of objects and tags.

- *Objects* are anonymous collections of *Tags*.
- *Objects* have no inherent meaning or owner.
- *Objects* are never deleted.
- *Tags* can be assigned any value, or none.
- *Tags* are protected by *Permissions*, but *Objects* are openly writeable.
- *Tags* can be modified or deleted by anyone with the appropriate *Permissions*.
- *Tags* can be categorized and collected into *Namespaces*.

[a]`http://www.fluidinfo.com`

Use Cases

The Einstein Toolkit [3] is a community-driven component framework for relativistic astrophysics, which is based on the Cactus Framework [1]. Using Fluidinfo's flexible data model, members of the community can easily create objects for their codes and annotate these with the following types of standard data: (1) standard metadata such as authors, languages, etc.; (2) datasets generated by the code; (3) dependencies on other codes. Meanwhile, a student using the codes could annotate the same objects with their own functionality rating. These two sets of tags, and others, can coexist without interfering with each other, and without anyone having to foresee the additional types of data which might be added.

Fluidinfo Query Language

Fluidinfo includes a simple query language to allow users to search the datastore for specific tags and tag-values. There are five basic types of queries in Fluidinfo's query language.

Presence queries check only for the presence of a tag on an object.
```
has eric/seen
```

Numeric queries search for tags that have a specific value.
```
eric/rating >= 4
```

Textual queries attempt to match the query text against the text contents of a tag.
```
eric/opinion matches "good"
```

Set queries attempt to match a string to the elements of a `set of strings`.
```
cactuscode.org/authors contains "Eric L. Seidel"
```

Logical queries combine the above types using the `(`, `)`, `and`, `or`, and `except` operators.
```
(mike/rating > 8 or imdb.com/rating > 7) except has eric/seen
```

Core Capabilities

- Code or data providers create their own objects, or if they don't exist others can create them.
- Virtual organizations, such as the Einstein Toolkit Consortium, can provide structured and reliable metadata, such as basic information about code components.
- Anyone else can provide complementary, or alternative, metadata to the same objects which does not interfere with other metadata sets.
- Tools can easily be constructed that allow searching over different metadata sets, e.g. combining data from the Einstein Toolkit Consortium and a trusted colleague.

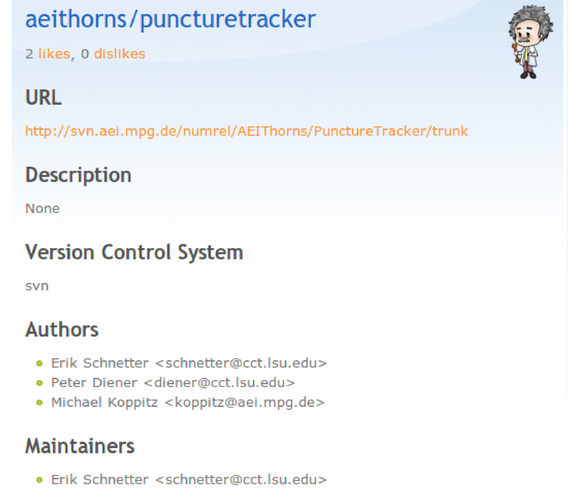

Figure: A prototype of a web application that dynamically displays thorn metadata based on the tags stored in Fluidinfo. The Einstein logo in the top-right corner indicates that this thorn is part of the Einstein Toolkit.

References

[1] Cactus Computational Toolkit, URL `http://www.cactuscode.org`.

[2] Dublin Core Metadata Initiative, URL `http://dublincore.org`.

[3] The Einstein Toolkit, URL `http://www.einsteintoolkit.org`.

[4] Fluidinfo, URL `http://www.fluidinfo.com`.

Acknowledgements

This work was supported by the Blue Waters Undergraduate Petascale Internship and Fluidinfo, Inc. We acknowledge Nicholas J. Radcliffe, who created `http://abouttag.com` for generating visuals of Fluidinfo objects. Eric Seidel is currently working for Fluidinfo as a summer intern.

NSF LSU CENTER FOR COMPUTATION & TECHNOLOGY the City College of New York