Beyond Galaxy Portable workflow and tool descriptions with the CWL*

*common workflow language

Michael R. Crusoe 2015-07-08. GCC 2015, Norwich

#CommonWFL

Michael R. Crusoe

Staff Software Engineer

The Lab for Data Intensive Biology

C. Titus Brown's k-h-mer project,
now at University of California, Davis.

@biocrusoe

Audience for this talk

Tool & workflow authors

Tool & workflow users

Platform developers

What’s the problem?

Need interoperable description of how to invoke non-interactive POSIX tools and how to describe the data flow between such tools.

Use cases include

1. Generate GUIs for command line tools
2. Tool authors ship their own tool descriptions
3. Remix workflows & run on the platform of your choice

HOW STANDARDS PROLIFERATE:

(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)

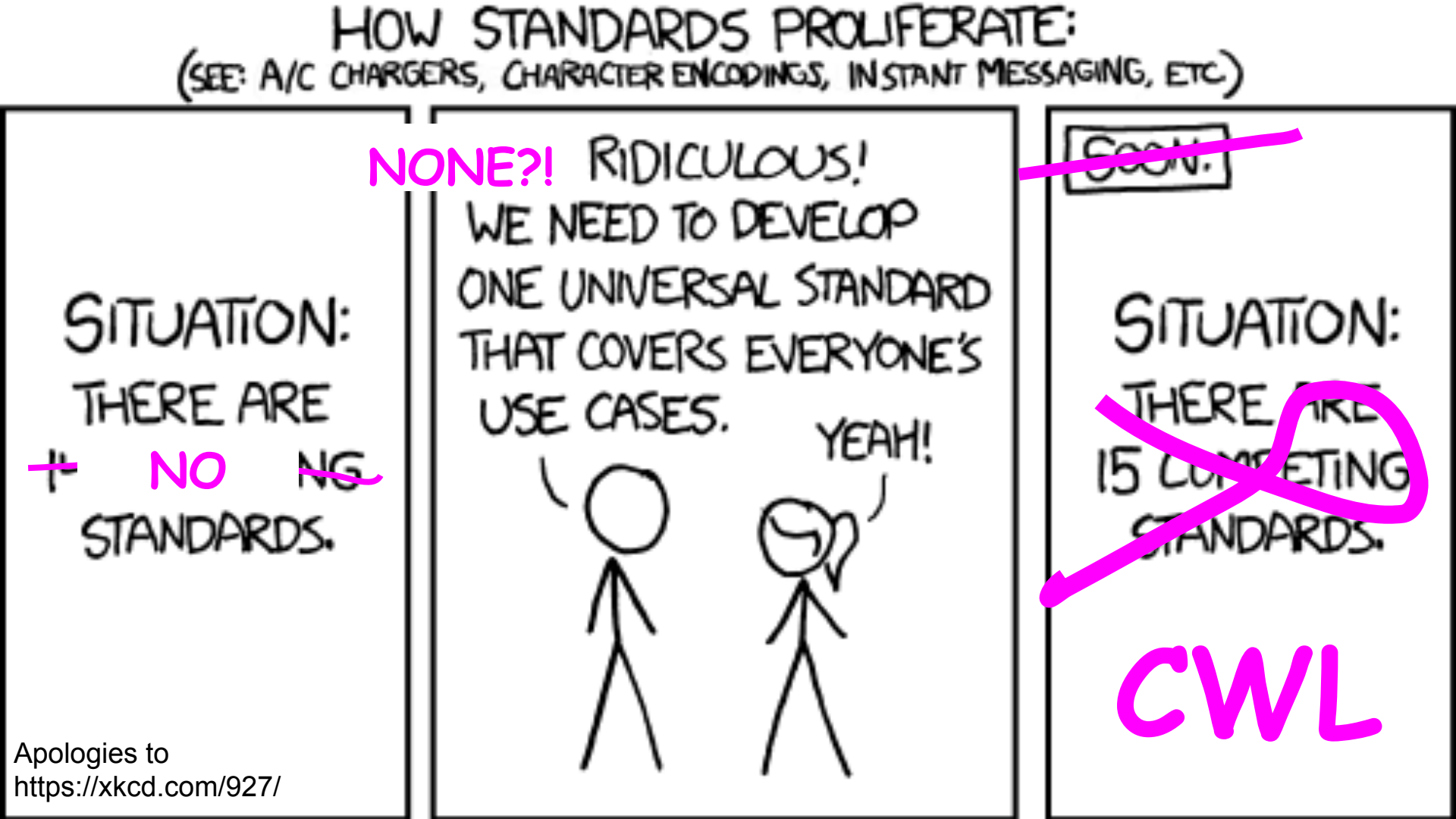
SITUATION: THERE ARE 14 COMPETING STANDARDS.

14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.

SOON:

SITUATION: THERE ARE 15 COMPETING STANDARDS.

https://xkcd.com/927/


HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)
SITUATION:
THERE ARE
NO
STANDARDS.
NONE?!
RIDICULOUS!
WE NEED TO DEVELOP
ONE UNIVERSAL STANDARD
THAT COVERS EVERYONE'S
USE CASES.
YEAH!
SOON:
SITUATION:
CWL
Apologies to
https://xkcd.com/927/

Other tool description approaches

Galaxy's "tool config file": leaks Galaxy internals

Taverna's tool service: format not documented

EMBOSS's ACD: only for EMBOSS style tools (great docs though)

iPlant's DiscoveryEnvironment: GUI only; no import/export

Other workflow descr. approaches

- Best model we found: Workflow4Ever

http://www.wf4ever-project.org/

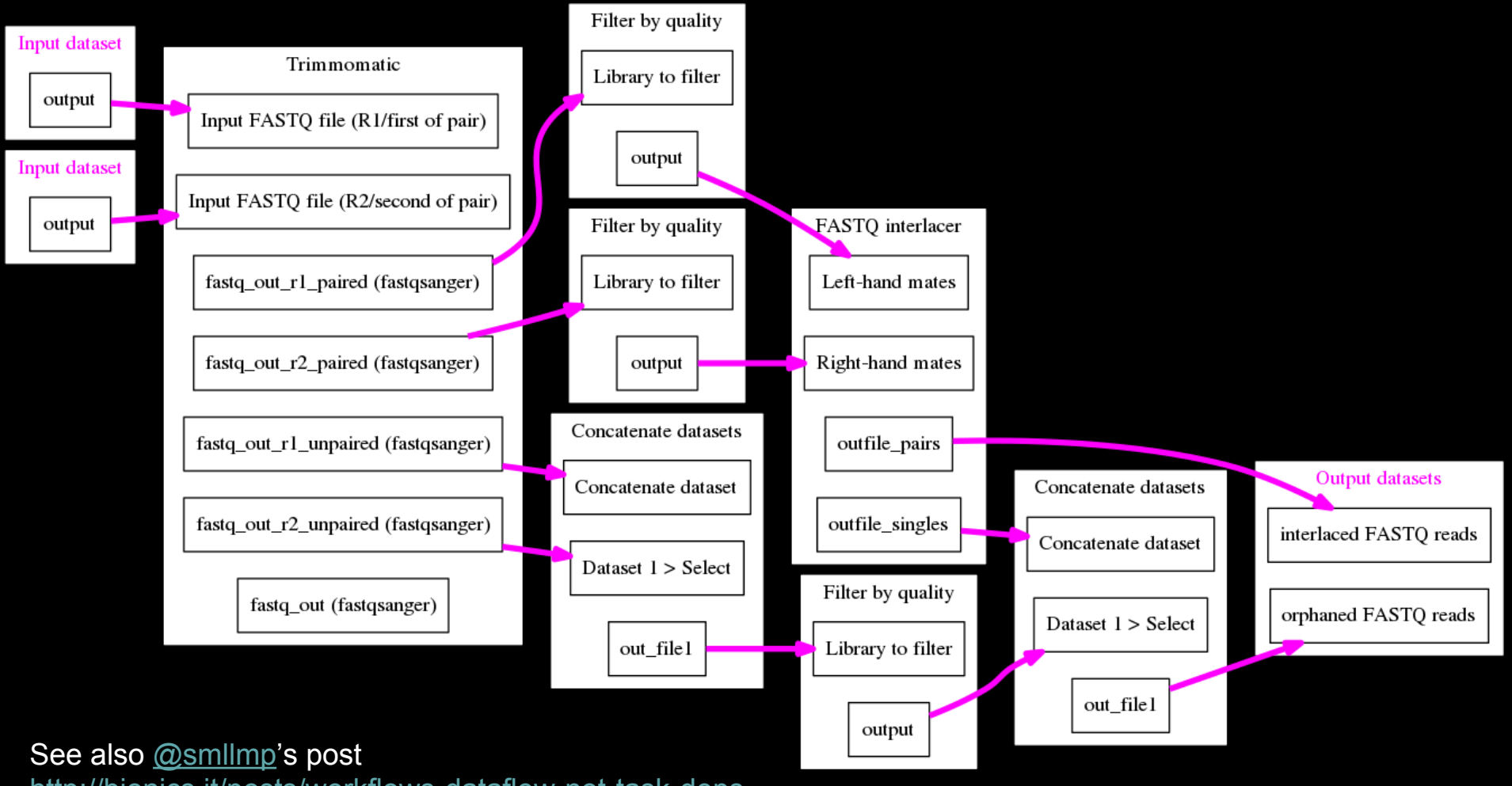
- All other descriptions were heavily tied to a single implementation

Features: (1) YAML: No XML!

```
#!/usr/bin/env cwl-runner

class: CommandLineTool
description: "Sort lines using the `sort` command"

inputs:
  - id: "#reverse"
    type: boolean
    inputBinding:
      position: 1
      prefix: "--reverse"
  - id: "#input"
    type: File
    inputBinding:
      position: 2

outputs:
```

Features

- Workflow linkages follow the data
- Data can be files on disk or streamed
- Scatter/gather for parallelism
- Built-in docker support



See also @smllmp's post
http://bionics.it/posts/workflows-dataflow-not-task-deps

More features

- Expression engines are pluggable (javascript, Python/cheetah, ...)
- Extensible
- Uses a linked data format that can be layered on top of
  - EDAM for metadata

Improving Galaxy - Auto Format

- Added support for auto_format="True" on Galaxy tools.
  - Causes outputs for tools to be sniffed after the job is complete.
  - Important applications for data source tools.

Improving Galaxy - EDAM Support

- All built-in Galaxy datatypes have been annotated with EDAM types (in dev branch)
- The Common Workflow Language has agreed to use Galaxy short identifiers as aliases for EDAM numeric IDs.
- Important applications such the ELIXER tool registry (slides by Olivia Doppelt-Azeroual et. al. http://bit.ly/GCC_ReGaTE).

Refactoring Tool Concept from Storage

- Abstracted Galaxy's generic tool representation from XML parsing.
- Abstractions make it possible to support other tool formats.
  - Checkout a YAML representation of tools which can be configured to run Galaxy today https://github.com/galaxyproject/galaxy/blob/dev/test/functional/tools/simple_constructs.yml

```
id: simple_constructs_y
name: simple_constructs_y
version: 1.0
command:
  >
    echo "$booltest"  >> $out_file1;
    echo "$inttest"   >> $out_file1;
    echo "$floattest" >> $out_file1;
    cat "$simp_file"   >> $out_file1;
    cat "$more_files[0].nestinput"   >> $out_file1;
    echo "$p1.p1val"  >> $out_file1;
inputs:
- name: booltest
  type: boolean
  truevalue: booltrue
  falsevalue: boolfalse
  checked: false
- name: inttest
  type: integer
  value: 1
- name: floattest
  type: float
  value: 1.0
- name: simp_file
  type: data
```

What should you do about it?

Tool & workflow authors: test the spec out, did we miss anything?

Tool & workflow users: ask your platform to support importing and exporting CWL descriptions

Platform developers: implement CWL support; hack on the reference implementation.

Next steps

GUI for workflow viewing / edit

Additional implementations by bioinformatics platforms

Packaged software requirements using Debian / PyPI / et cetera

How to work with the CWL group

- Draft 2, link to GitHub repos & mailing lists at http://common-workflow-language.github.io/
- (Every 2-3 weeks we have a Google Hangout)
- Birds of a Feather session **tonight**, 18:00 Franklin Room, JICCC
- Birds of a Feather session Saturday afternoon at BOSC (Dublin)

Thanks!

Curoverse, Peter Amstutz

Seven Bridges Genomics, Nebojša Tijanić

Galaxy Project, Pennsylvania State University, John Chilton

Institut Pasteur, Hervé Ménager

BioDatomics, Maxim Mikheev

University of Manchester, Stian Soiland-Reyes

@biocrusoe funded via the U.S. National Human Genome Research Institute of the National Institutes of Health under Award Number R01HG007513, then at Michigan State University

Institut Pasteur