Google Summer of Code – Graphical editor of XML files with Blockly

DataReview, 5th Sept 2016

Matthieu Muffato
Ensembl Compara / eHive

EMBL-EBI

wellcome trust sanger institute

eHive *PipeConfig* files

Perl modules:

- Prone to syntax errors (missing comma, missing semicolon, etc)
- No validation tool
- Fake keywords like WHEN, INPUT_PLUS to extend the language
- Can do naughty things like

```
if (($self->o('division') !~ /^#:subst/)
    and (my $tl = $self->default_options()->{'taxlevels_'.$self->o('division')})) {
       if (stringify($self->default_options()->{'taxlevels'})
           eq stringify($self->o('taxlevels'))) {
               $analyses_by_name->{'group_genomes_under_taxa'}->{'-parameters'}
                                                                ->{'taxlevels'} =
$tl;
       }
}
```

Alternatives to Perl

Considered XML and JSON with these goals:

1) There must be a schema specification and a validator

2) We want an editor that understands the schema

- Investigated XML and XSD or RelaxNG (RNG)
- Investigated JSON

Blockly

- Library developed by Google

- Programming editor using visual blocks (used to teach programming)

- Very versatile: can define your own block and rules, has an API / hooks to export a diagram to code

https://studio.code.org/s/course1/stage/18/puzzle/
https://blockly-demo.appspot.com/static/demos/code/
http://ens-lg4.github.io/MenulyJSON/
http://muffato.github.io/eHive-Blockly/

Relax NG

Schema definition for XML files

- The attributes of each node element
- Where free text can be placed
- How elements can be nested

- Used by the LRG project !

RNG / XML example

RNG

```
<?xml version="1.0" encoding="UTF-8"?>
<grammar xmlns="http://relaxng.org/ns/structure/1.0">
  <start>
    <element name="contact">
      <element name="name">
        <text/>
      </element>
      <element name="phone_number">
        <text/>
      </element>
      <element name="email">
        <element name="username">
          <text/>
        </element>
        <element name="hostname
          <text/>
        </element>
      </element>
    </element>
  </start>
</grammar>
```

XML

```
<contact>
  <name>Matthieu</name>
  <phone_number>01223494631</phone_number>
  <email>
    <username>muffato</username>
    <hostname>ebi.ac.uk</hostname>
  </email>
</contact>
```

Block patterns in RNG

- <optional> <...A...> </optional> **A?**
- <oneOrMore> <...A...> </oneOrMore> **A+**
- <zeroOrMore> <...A...> </zeroOrMore> **A***
- <choice> <...A...> <...B...> … </choice> **A|B|...**
- <interleave> <...A...> <...B...> … </interleave> **A & B & …**

- Can be nested: **(A | B+) & C* & (D? | (E+ & A))**

valid patterns: A, DBBCC, AAEE

GSoC project

- Parse the RNG file and define Blockly blocks
- Define Blockly rules for the block connections
- Write an extra validator for the rules that could not be modelled in Blockly
- Code to export a diagram to XML
- Code to import an XML file to a diagram

Google
Summer of Code

Anuj Khandelwal @anujk14 (GitHub)

Demo time !

https://ensembl.github.io/XML-To-Blockly/

Conclusions

Some work left to do:

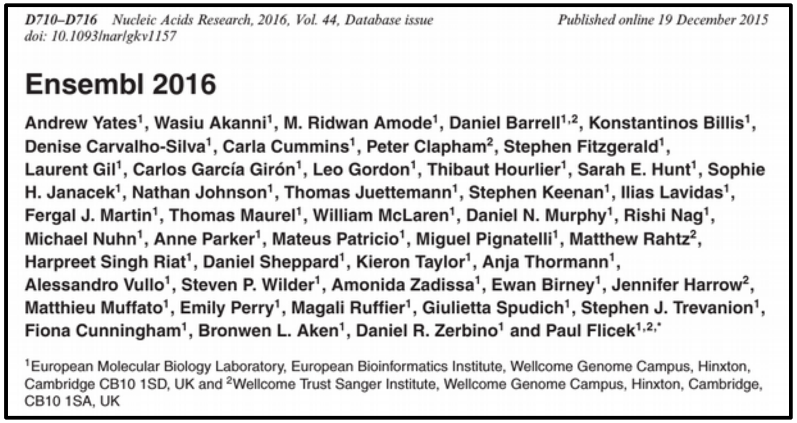
- Import of an XML file
- Interface improvements (zoom, collapse, etc)
- Scalability of the interface for large pipelines ??

Acknowledgements

Anuj

Google
Summer of Code

Leo



D710–D716 Nucleic Acids Research, 2016, Vol. 44, Database issue *Published online 19 December 2015*
doi: 10.1093/nar/gkv1157

Ensembl 2016

Andrew Yates[1], Wasiu Akanni[1], M. Ridwan Amode[1], Daniel Barrell[1,2], Konstantinos Billis[1], Denise Carvalho-Silva[1], Carla Cummins[1], Peter Clapham[2], Stephen Fitzgerald[1], Laurent Gil[1], Carlos García Girón[1], Leo Gordon[1], Thibaut Hourlier[1], Sarah E. Hunt[1], Sophie H. Janacek[1], Nathan Johnson[1], Thomas Juettemann[1], Stephen Keenan[1], Ilias Lavidas[1], Fergal J. Martin[1], Thomas Maurel[1], William McLaren[1], Daniel N. Murphy[1], Rishi Nag[1], Michael Nuhn[1], Anne Parker[1], Mateus Patricio[1], Miguel Pignatelli[1], Matthew Rahtz[2], Harpreet Singh Riat[1], Daniel Sheppard[1], Kieron Taylor[1], Anja Thormann[1], Alessandro Vullo[1], Steven P. Wilder[1], Amonida Zadissa[1], Ewan Birney[1], Jennifer Harrow[2], Matthieu Muffato[1], Emily Perry[1], Magali Ruffier[1], Giulietta Spudich[1], Stephen J. Trevanion[1], Fiona Cunningham[1], Bronwen L. Aken[1], Daniel R. Zerbino[1] and Paul Flicek[1,2,*]

[1] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK and [2] Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

Funding:

wellcome trust

National Human Genome Research Institute

BLUEPRINT epigenome

EMBL

BBSRC bioscience for the future

EURATRANS

Co-funded by the European Union