Hacking groups

BioHackathon 2024

DBCLS
BioHackathon

Use this slide as a template

Group topic

Participants

- Toshiaki, …

Description

-

← Put your names here to participate in
Use the bold font for group lead

← Describe objectives, requirements etc.
(Group leader is responsible for this and presentation)

We will start presentations and reviews of the hacking projects at 11:00

Suggested target domains

- [R1] Multi-omics analysis on human genotype to phenotype that includes genomic, transcriptomic, epigenomic, proteomic, protein structures, and biochemical data.
- [R2] Automated data analysis of other organisms including phylogenetic compositions, gene annotations, pathways, and growth conditions.
- [R3] Data-driven interdisciplinary studies in public health, environment, agriculture, food, energy, and other fields utilizing knowledge graphs.
- [R4] Facilitating knowledge discovery and biological analysis from databases and literature, especially utilizing large language models.

R1

Human genotype to phenotype

Genome variation

Participants

- **Yosuke**, Toshiaki, Maxat, Toyoyuki, Yuki, Nobutaka, Hirokazu, Tsuyoshi, Dorothy, Pitiporn (Sam), Mayumi, Núria, Shuichi, Kentaro (Yamaken), David, Takatomo, Hiroyuki Mishima, Tazro

Motivation

The analysis of the human genome has been flooded with data due to the widespread use of sequencers. The simple variations such as SNV and indel are being integrated with TogoVar, but structural variation has not yet even been standardized. Meanwhile, the pangenome graphs has emerged as a powerful tool for integrating multiple haplotypes. During this hackathon, we would like to discuss how to handle this heterogeneous data in a unified manner.

Description (Write down your proposal here.)

- Genome graph - new repository -> move to new project
- SV, CNV, STR, …
- Clinical variants - MGeND
- https://www.ga4gh.org/product/variatioepresentation/ (are you also interested?👍)
- Workflow (Tool Pipeline) n-r
- Genome Variation Ontology http://genome-variation.org/resource/gvo
- Development of consensus pangenome graph tool

Join #genome-variation channel on slack.

Pangenome Graphs Database (PGD)

Participants
- Toshiaki, Yosuke, Maxat, Robert (remotely), Toyoyuki

Description
- Collect existing pangenomes into one repository
  - Do a survey on the papers published so far
    - Human: HPRC, Chinese, Arab, JaSaPaGe,
    - Primate:
    - Plants:
  - Grep existing BioProjects with the term /pangenome/
    - => 609 entries
  - GitHub repository for the draft version:
    - https://github.com/JaSaPaGe/pangenome-graphs
- Define metadata in JSON (and turn it into JSON-LD by adding @context later)
  - Target (population) - do we also include non-human pangraphs?
  - (Number of) samples (haplotypes?)
    - Links to raw data and assembled haplotype sequences (e.g., SRA)
  - Availability
    - Download link - do we copy the graph data into our database? to GFA
  - License
    - Need to contact to the authors for data retrieval?
    - Requirement of IRB (institutional review board) approval
  - Method
    - Workflow and tools used to create the graph - link to the repository?
  - Authors
    - Contact information
  - Reference
    - Published paper on the graph
  - Version
    - Published date, Updated date and revisions
- A Website with a SPARQL endpoint
  - pgd
- Analysis environment
  - Should be replicated in cloud environments and on-premise systems
- Submit the database to a journal (at least to the BioHackrXiv)

Integrating facial analysis into PubCaseFinder

Slack Channel: #pubcasefinder_gestaltmatcher

Participants
- **Tzung-Chen Hsieh**
- Hiroyuki Mishima
- Toyofumi Fujiwara (remote)
- Marlon Aldair Arciniega Sanchez
- Atsuko Yamaguchi (interested)

Description
- PubCaseFinder (https://pubcasefinder.dbcls.jp/), the framework to search for disorder/gene/patients by Human Phenotype Ontology (HPO) analysis.
- GestaltMatcher Database (GMDB, https://db.gestaltmatcher.org/), the database containing ~10,000 facial images with rare disorders.
- To PubCaseFinder, implement functionalities to link GMDB and perform diagnosis assistance using facial photos.
- Input: facial image and HPO terms
- Output: a list of suggested disorder/genes/patients. Additionally, show the links to the photo in GMDB.
- Test data: GMDB test set and published Japanese patients from internet.

HPO suggest

Participants
- **Marlon Aldair Arciniega Sanchez**
- Toyofumi Fujiwara (remote)
- Atsuko Yamaguchi
- Orion Buske (remote)
- Andrea (maybe)
- Yosuke Kawai [interested]
- Toyoyuki Takada [interested]

Description
 Objectives
 - Given (one or more) HPO terms, suggest one or more HPO terms based on the log of PubcaseFinder queries
 - Analyze PubCaseFinder queries for any biases or usage patterns to better understand users
   - bias in branches of HPO being searched for, JP vs EN, IP/geography, terms in particular order

Methods
- Data cleaning (deduplicate sequential queries from same user?)
    - Use data, time, or IPs related to each query
- Create a matrix of the co-occurrences between HPO terms
    - Calculate conditional probabilities given the frequency of each HPO and its combinations.
- Measure performance against searches in other time period

Hidden-Rad ontology

Participants

- Key-Sun Choi (if others!) Andrea: interested but not sure if can participate. `Hikaru (interested)`

Description

- Task is to give an ontology for the radiology disease decision process about
  - findings, anatomical location, impression, and checklist to confirm the impression (disease)
  - In the environment of Radiology report based on MIMIC Chest-Xray data
- Checklist will be a clue for explaining why such disease impression was made, but usually not written in the radiology report.
- Now a collocation of data for such checklist confirmation has been made from the patient data in MIMIC by experts.
- https://sites.google.com/view/ntcir-18-hidden-rad/hidden-rad
  - To generate a report to include the explanation why such impression is made in the radiology report,
    - for input from MIMIC.
  - training data is generated by LLM based on experts' checklist confirmation by crowdsourcing and corrected by experts.
- Ontology schema consists of the
  - base ontologies: FMA, RadLex, DOID, MONDO
  - properties from RadGraph
  - specially required schema for checklists

Connecting healthcare data

Participants

- Andrea (if others are interested), Kiyoko (interested), Pitiporn(intersted)
- Evan (interested), Key-Sun (interested), Hikaru (interested), Núria (interested), Chihiro (interested), Toshiaki, Chang (interested)

→ can we connect this to clinical trials, having CT as a starting point and expanding from it?

Description

- Make a map of what connections exists between data/ontologies for patient data and molecular (or environmental) data/ontologies
- e.g.: From symptoms, to diagnosis, genetic basis, pathways, to chemicals and pollution and environment.
- Can we make a chart?
- Looks like Med2RDF is very related

Annotations of clinical trials

Participants

- **Thomas Liener,** Jerven Bolleman, Claude Nanjo, Núria (Andrea possibly interested), … ?, Dani F(maybe can help with the model), Yuka (interested), Tore (interested)
- Evan (interested - we worked with community to get CTO updated/modernized - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9389640/ .. PubChem has linking to clinical trials and we did in PubChemRDF but we did not release it for some reason .. I am asking why), Chunlei (interested)

→ can we connect this to “connecting healthcare data”, looking on how we can expand from basic elements in a CT? Drug, indication, symptoms, phenotypes to, for instance, what pathways are clinical trials about? What environmental factors link to a disease?

Description

- Annotating clinical trial data from https://clinicaltrials.gov/ with ontology terms
  - *How deep? You have MESH keywords, but also I/E criteria (quite convoluted) sites, protocol aspects… Is there even an ontology for to annotate protocols (maybe clinops? / USDM?)*
  - Entity Recognition necessary? LLM? Using existing resources (FHIR, existing MESH annotations for indications)?
- Building a (simple?) semantic model for clinical trials
- Linking/connecting clinical trials to other resources (Uniprot, pubchem?)

Brainstorming gdoc here and brainstorming slack #clinical_trials

Visualization for cohort data

Participants

- Akio Nagano, Yosuke Kawai [interested], Chihiro (interested), Michel (interested), chang

Description

- A tentative plan for visualizing cohort data
- *In which direction?*
  - I'm going to try a (maybe) slightly new way of visualizing information.
  - Cohort participants are represented as dots.
  - The dots will change shape depending on the visualization you're trying to achieve.
  - For example, if you're representing a histogram, the dots representing cohort participants will move to the bin they belong to and become part of the bars.

R2

Other organisms

Viral phylogenomics

Participants

- **Russell**, Yosuke, David (interested)

e.g. : https://ggdc.dsmz.de/victor.php

Problem : Species trees are usually built using sets of universal marker genes. Viruses don't have universal genes!

Proposal : Cluster gene trees by topology, build species trees for taxonomic groups with compatible gene trees.

Motivation : Species trees are the starting point for studying recombination among viruses and their hosts, testing models of species concepts in viruses, illuminating the origin of cellular and viral life, and many other things.

Dataset : IMG/VR (https://img.jgi.doe.gov/cgi-bin/vr/main.cgi) is the largest collection of viral genomes, with 5,576,197 genomes and MAGs in 2,917,521 vOTUs spanning all clades of the viral world.

Workflow : 木槌 (kizuchi) (https://github.com/ryneches/kizuchi/) uses prodigal-gv, hmmer, mafft, trimal, and fasttree to generate gene trees.

Cultivation media & phenotypic traits

Participants

- Julia, Shuichi, Risa, Kohei, Yoko, Tatsuya, (Erick: curious), Susumu (interested)

Description

- Sharing media information between MediaDive and TogoMedium
  - Understand the structure of each terminology
  - Align terminologies manually and automartically
  - Expanding and applying the cultivation media ontology
  - Developing an exchange format for media between the two platforms
- Calculating similarity between media
- Integrating more information on media design
- Cleansing phenotypic trait data and test data

- If time allows it: strategies and prototypes for AI-prediction of cultivation media

TOGO MEDIUM MediaDive

R3

Broader life sciences

GLYCO and all things sweet!

Participants

- **Kiyoko**, Evan, Issaku, Nathan, Akihiro, Masaaki, Masae, (Núria interested), Dani (could help here).

Description

- The structure of glycans lacks clear, distinguishable definitions, making it problematic to determine which structures should be considered as glycan data.
- Therefore, we will analyze the data from GlyTouCan, a glycan structure repository, to understand what structures are considered glycans and what structures are considered monosaccharides.
- We aim to discuss the results and establish rules for defining which structures should be classified as glycans.
- GlyCosmos development:
  - Archetypes and subsumption (Akihiro with help from Masaaki)
  - Motifs (Masae)
  - GlyCosmos RDF for RDF Portal (Masae)
- Development of tools
  - Update GlycanBuilder2, GlycanFormatConverter, wurcs2pic, etc.

Human Glycome Atlas (HGA)

Participants

- **Kiyoko**, Achille, Ruwan, Hannah

Description

- Evaluate various infrastructure components
  - QLever
  - GRASP
  - UniProt
  - Others?
- Try to load GlyCosmos RDF into QLever to assess its performance

PubChem ⇔ Nikkaji Alignment

Participants

- **Yuka**, Evan, Tatsuya, Issaku

Description

- Update Data in PubChem originated from Nikkaji
  - Remove duplicate entries (same CID, different SIDs) in PubChem
  - Remove inconsistencies between Nikkaji/Pubchem (Nikkaji ver 2018) and Nikkaji RDF (Nikkaji ver 2022)
  - Set up the procedure for finding inconsistencies and upload new Nikkaji entries to PubChem

Plant Breeding Ontology(PBO)

Participants

- **Erick Antezana**, Hiromi Kajiya-Kanegae, Shuichi, Wasin Poncheewin, Núria, Akio Nagano

Description

- Update the current version of PBO (OBO and RDF)
  - add new terms
  - refine some definitions
  - add Japanese translations
  - add new « categories » = hierarchy
- Review & update the support scripts (Python)
- Load PBO (in RDF) into a triple store
- Generate a few sample queries
  - on PBO
  - combining other resources (federation?)
- Explore new oportunities
- Publication
  - update the draft
  - japanese characters as images (fix)
- We need a nice ontology image/logo ~~← CALL FOR ARTISTS!~~ **FOUND!**

Japanese Food Ontology

Participants

- Chihiro, Tatsuya, Kiyoko (interested), Erick A. (interested), Shuichi (interested), Risa (interested), Núria (interested), Julia (interested), Susumu (interested)

Description

- Integration Japanese food ontology using another Japanese food resources
  - based on FGNHNS (https://bioportal.bioontology.org/ontologies/FGNHNS)
  - addition of food composition data (MEXT)
  - addition of standard food name data (MIC)
- Consideration
  - addition of allergen information
  - relation of crop information
  - relation of FoodOn

BH24 Wikiblitz (fun and sidetopic)

Participants

- Andra, Yasunori, Shuya, Russell, Michael、Tore

What is a Wikiblitz?

A **Wikiblitz** combines **Wikidata/Commons** with a **Bioblitz**:

- A Bioblitz is a communal effort to record as many species as possible within a specific location and time.

Why Participate?

- Your observations, under an open license, can be reused.
- Using Wikidata, we link these observations to the semantic web.
- You might even discover a species not yet observed!

Join Us!

- **Slack:** #wikiblitz
- **iNaturalist:** Biohackathon 2024 Project

Let's explore and contribute together! (Maybe interesting? https://www.earthmetabolome.org/)

R4

Data analysis and methods

Hindsight/best practices

Participants

- Jerven, Evan, Yoko, Erick, Andrea, Andra…, Yasunori, Michel, Julia, Thomas
- Jose (interested), Arto (interested), Takatomo (interested), Shuichi (interested), Chunlei (interested), Risa (interested)

Description

- UniProt, PubChem, Rhea
  - We did some stuff, what do we regret, what do we want to improve
  - Can we “fix” it in spec compatible ways (e.g. owl:equivalentClass)
- Advice for the next gen
  - Query optimizer friendly
  - Human friendly SPARQL, RDF and identifiers
  - Long term data preservation
  - Multi-Language support
- Input from data integrators
  - What do they love/hate?
  - What is best way to improve interoperability of RDF data sets?

docker run -p 5000:5000 gm-api

Getting shapes from large RDF inputs

Participants

- Dani, Yasunori Yamamoto, Jose Labra, Andra Waagmeester, Jerven
- Evan (interested)
- Gos Micklem

Description:

We aim to automatically extract RDF shapes (ShEx, SHACL) from large data sources. To address scalability challenges, we've developed a solution that involves splitting the input source into manageable slices and then merging the resulting schemas. However, 1) this is just one approach, and 2) we need to enhance the subsetting process to ensure the subgraphs are as complete as possible while remaining manageable by commodity hardware.

We would appreciate assistance with:

- Developing subsetting strategies
- Suggesting parallelization techniques
- Hands-on support for implementation

Visualize sheXer results

Participants
- Dani
- Kozo
- Andra
- Gos
- Jose (interested)
- Toshiaki

Introducing sheXer: Automate Your RDF Schema Inference!
- **What is sheXer?**

A Python library that automatically infers RDF schemas.
- **Current Features:**
  - Outputs **ShEx**, **SHACL**, and **PlantUML** visualizations.
- **Our Goal:**
  - Enhance schema visualizations with new and diverse visualization backends.

Slack: #shexer

Using (discovered) schema

Participants
- Dani, Jerven, Jose Labra, Núria, Yasunori, Andra
- Evan (interested), Chunlei (interested)
- Gos, Toshiaki

Description
- We can use shexer or void-generator or rdfdoc to discover the schema of data.
- With the schema we can
  - generate code
    - generate an RDF-config model stab file. (hopefully automatically name variables based on Class/Property labels (rdfs:label / rdfs:comment))
  - validate/generate sparql using rudof
  - link examples to the schema
  - improve query auto complete

LLM-SPARQL

We propose to develop an **LLM-assisted SPARQL query answering system**

- schema-informed *in context learning* by LLM
- corrective SPARQL query generation
- evaluation over human and AI generated benchmarks

Tasks:

1. **SPARQL query benchmark** (human and AI generated)
2. **LLM-framework** (llama-index)
   a. identify schema-relevant info from user query
      i. NLP, schema extraction & store, schema mapping, ontology reasoning, graph analysis
   b. generate SPARQL query
      i. baseline LLMs: local (llama3.1), cloud (GPT-4o)
      ii. constrained sampling (syntax-directed token selection)
      iii. fine-tune LLM
      iv. (train a new generator via stable diffusion)
   c. validate and iteratively generate SPARQL query
      i. analyze syntax and semantics
      ii. suggest improvements
3. **Evaluation**

Project Slidedeck

Participants

- Michel Dumontier
- Jerven Bolleman
- Andra Waagmeester
- Hikaru Nagazumi
- David Steinberg
- Chang Sun
- Arto Bendiken
- Yasunori Yamamoto [interested]
- Claude Nanjo
- Eric Prud'hommeaux
- Jose Labra
- Gos Micklem
- Dani Fernández
- Julia (interested)
- Chihiro (interested)
- Chunlei
- Shuichi [interested]
- Toshiaki

SPARQL - Schema conversions

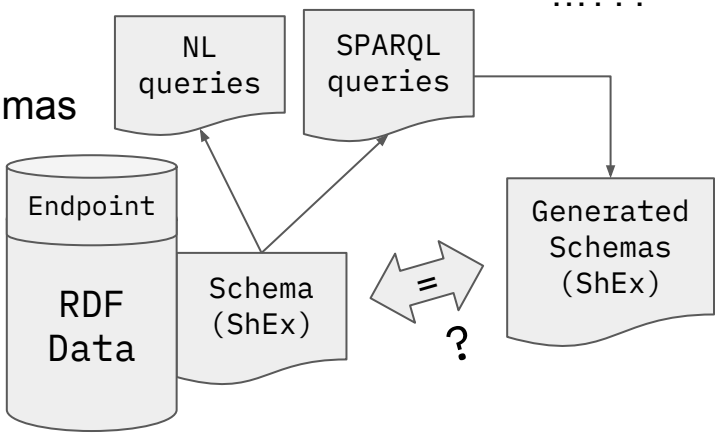
Participants:
Jose Labra,
Hikaru Nagazumi,
Eric Prud'hommeaux,
Claude Nanjo
Andra
Yasunori Yamamoto
Dani
Gos
…???

Building blocks identified as part of the LLM SPARQL project

- Challenge 1: ShEx → NL Question + SPARQL :
  - To compare with the other direction: NL Question → SPARQL
  - SPARQL query benchmark
- Challenge 2: SPARQL → ShEx:
  - Goal: Schema extraction
- Challenge 3: Compare between ShEx schemas
  - Goal: Schema mapping
- Challenge 4: Visualize ShEx schemas
  - Goal: Help humans understand the schemas

Slack channel:

#sparql_schema_conversions



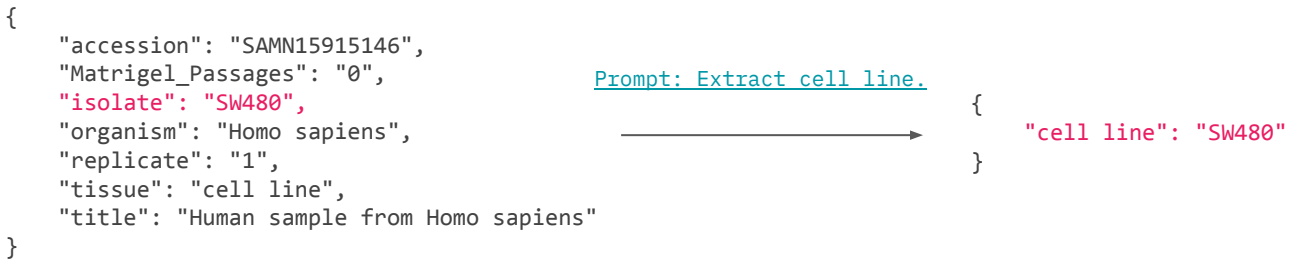
LLM-assisted BioSample curation

Participants

- **Shuya**, Tazro, Shinya, Zhaonan, Yuki, Takatomo, Shuichi, Susumu

Description

- Improve the quality of metadata registered in the BioSample database using LLMs
  - Metadata of BioSample is very heterogeneous and hard to interpret algorithmically
  - Extract phrases to be mapped to ontology terms
  - For evaluation of the task, create a testset manually
    - Complete the testset

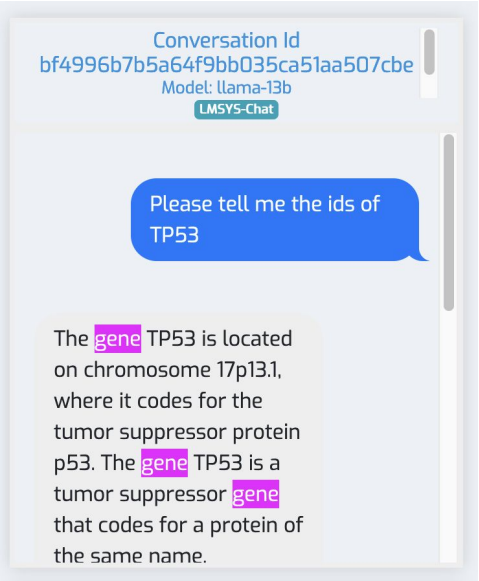


Characterize Biology use of LLMs

The interaction with LLMs presents a new way of using computers and should be studied directly.

WildChat dataset includes 1m real-world usages of ChatGPT including many biological questions. Let's find out what people used it for (and if it was any good!)

- Find subsets of WildChat dataset relevant to bioinformatics/biology
- Benchmark usage from other LLMs
- Summarize the usage (using LLM and hand curation)
- Review the quality of responses
- Generate synthetic ChatGPT conversations using WildChat model



Participants:
David
Hirokazu
Tazro [interested]
Toshiaki [interested]
Susumu [interested]
Pitiporn(Sam) [Interested]

Ruby coding with help of LLMs

Participants

- **Naohisa**, Hiroyuki, Arto,

Description

- Trying to write Ruby code to do Bioinformatics tasks with the help of ChatGPT and LLMs
- Developing Ruby libraries/applications with LLM
  - BioRuby: Bioinformatics library for Ruby
  - …

UMAP all the APIs

Make a visual representation of metadata about DBCLS services

- Create a sparse vectorized representation of API/services
- Generate a dimensionally reduced visualization of the services
- Provide an interactive interface for accessing underlying services
- Characterize clusters

Participants:

- David

Data quality

Participants

- Andrea (if others!), Kiyoko (interested), Yasunori
- Achille (Interested)
- Yuka (interested but not sure if I can participate)
- Erick (Interested)
- Evan (interested) , Jerven (Interested), Chunlei (interested)

Description

- How do you annotate a dataset quality, so that a consumer can evaluate if it is viable for a given purpose?
- Objectives of this project are to develop:
  - A set of properties for data quality labeling (e.g.: dataset coherence)
  - A set of metrics for such properties
  - Implementation for such properties

- Anybody interested, I have two elements to build on:
  - https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/data-quality-framework-eu-medicines-regulation_en.pdf (Data quality framework for EU medicines regulation - 2 years)
  - **(On the value of data - collaborative paper on data evaluation started some time ago)** **https://docs.google.com/document/d/1UrJ476aeQ1uKoFvqEzp-TiEdab2hXORy3hmwygpfLZQ/edit#heading=h.w78pl57gjfke**
  - Also related some work on RWD Metadata and data labeling
  - Fair Metrics? yummydata

Enhancement of the Bioresource Retrieval System using ChatGPT

Participants
- Tatsuya, Chihiro, Terue, Hiromi (interested)

Objectives:
- Implement the ChatGPT API in the bioresource retrieval system to enable fuzzy search capabilities.

Tasks:
- Integrate the ChatGPT API with the FileMaker system.
- Develop and refine effective prompts for the ChatGPT API.
- Test the fuzzy search results to ensure accuracy.
- Compare FileMaker+ChatGPT with other services, such as Dify, and NotebookLM.

Expected Results:
- Enable users to perform more intuitive and flexible queries.
- Enhance the overall user experience.
- Prevent failures in information retrieval due to user errors, such as typos or misspellings.
- Improve search accuracy, with a particular focus on enhancing recall.

Mass Spectrum Viewer

Participants

- Satoshi, Masaaki
- Erick (curious, developed IsotopIdent, https://bio.tools/isotopident)
- Evan (interested .. PubChem has an image generator for a set of peaks https://pubchem.ncbi.nlm.nih.gov/docs/imaging-services#section=Mass-spectrometry-image-svg)
- Nathan (have javascript spectrum/chromoatogram viewer)
  - See https://gptwiki.glyomics.org/gptwiki/TG001210 (annotated MS/MS),
  - https://gptwiki.glyomics.org/gptwiki/TG009582 (transition chrom.)

Description

- Developing some kind of viewer of mass spectrum data.
  - Heatmap

https://github.com/masspp

Workflow and Container helpdesk

Participants

- **Tomoya**, Kentaro(Yamaken), Pitiporn(Sam), Manabu, David, Naohisa, Michael (online), Arto (interested), Chihiro (interested)

Description

- Help others to develop their workflows
  - e.g., CWL, snakemake, nextflow, …
- Help others to use workflow-related technologies
  - Containers such as Docker, Singularity, Podman, …
  - Job Schedulers such as Slurm, GridEngine,…
- Develop and improve workflow ecosystems
  - e.g., executors, specifications, related tools, and workflows!

Help !!

- We want better name for our group !!

Slack channel: #workflows

Workflow and Container helpdesk

What we did:

- Report an issue to clarify a corner case in the spec of CWL v1.3
- Release a new version of shaft, an executor for CWL, to demonstrate new feature of CWL conformance test suite
  - Now users see the detailed results of each test category
  - Example: detailed result of the CommandLineTool category
  - Other CWL engines can easily implement the same feature with Custem GitHub Actions in the marketplace:
    - Run CWL conformance tests
    - Upload CWL conformance badges
- Report an issue when passing array input parameters via command line arguments in cwltool
- ddd

Additionals

Self-Introduction

Request for tutorials

- QLever: Arto, Gos, Toshiaki, Evan, Julia, Ruwan
- Neptune: Arto, Evan, Jerven
- RDF Portal: Andrea, Gos, Núria,
- RDF-config: Núria, Gos, Dani, Evan
- TogoVar:
- SPARQList
- TogoID: Evan, Chunlei
- PubCaseFinder: Andrea
- TogoDX
- TogoDB: Hiromi
- RDF-doctor
- SPARQL-proxy
- TogoStanza
- MetaStanza
- UmakaYummy
- PubAnnotation
- Grasp
- JSON2LD Mapper: Chunlei
- TogoWS
- TogoGenome: Sam
- Endpoint browser: Jerven, Chunlei
- TogoMedium
- D2RQ Mapper
- PubDictionaries: Andrea
- Allie
- SPANG
- Colil
- inMeXes
- Med2RDF: Andrea

Group photo

Let’s take a group photo in YUMORI before lunch!