CRIM

CENTRE DE
RECHERCHE INFORMATIQUE
DE MONTRÉAL

RDM-071 aka FRACS

FAIR Repository for Annotations, Corpora and Schemas

PRESENTED TO: CANARIE RDM Workshop
BY: Philippe Collard

January 23, 2019

crim.ca

PRINCIPAL PARTENAIRE FINANCIER Québec

BACKGROUND

- Projects in unstructured audio / video / image / text data valorisation
- Need supervised learning
- Need data that has been labeled / annotated
- Labeled dataset: created by a human who specifies what he sees or what he understands in the data
- Labeled data or annotations: ‘ground truth’ i.e. a training dataset that a researcher will use to train model with ML algorithms

BACKGROUND

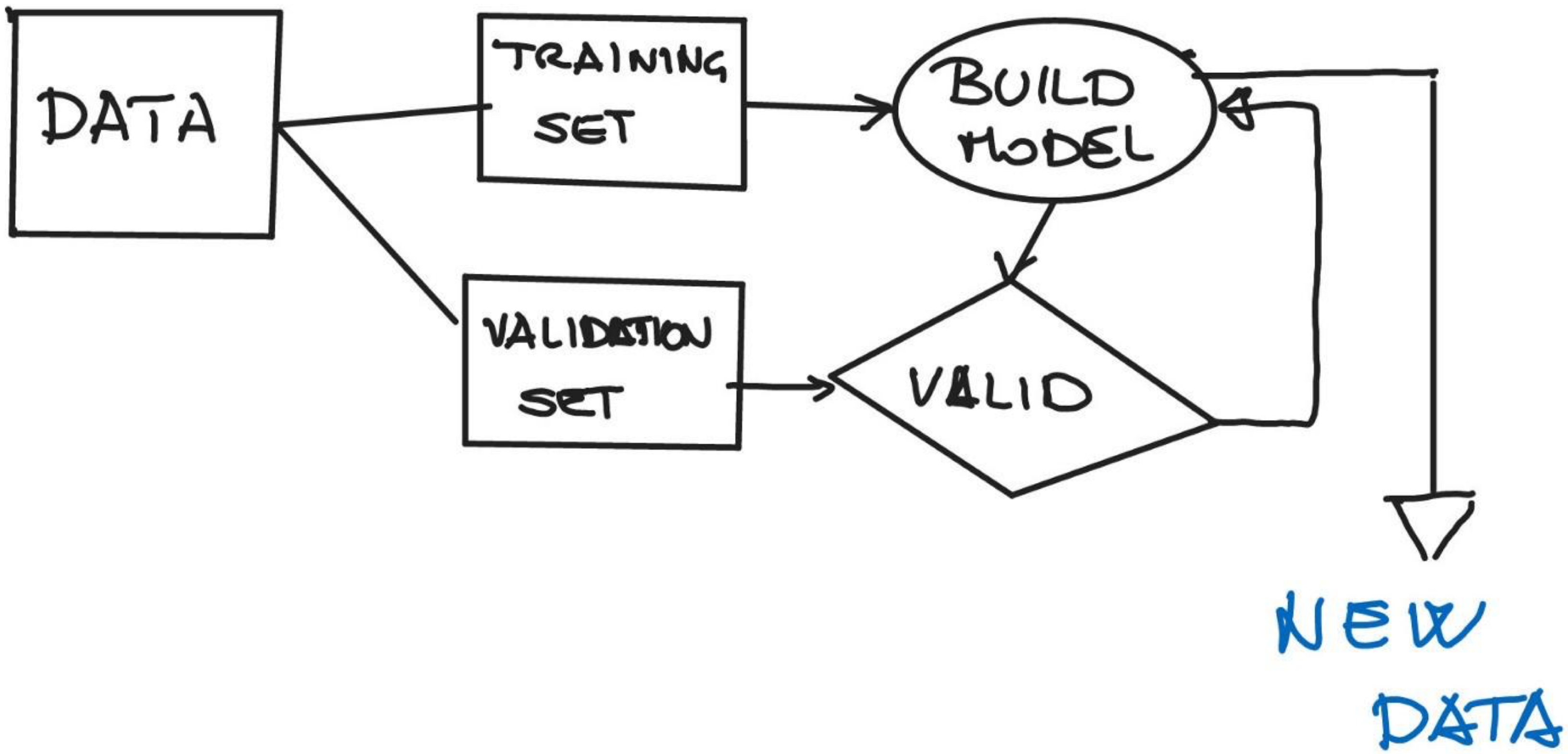
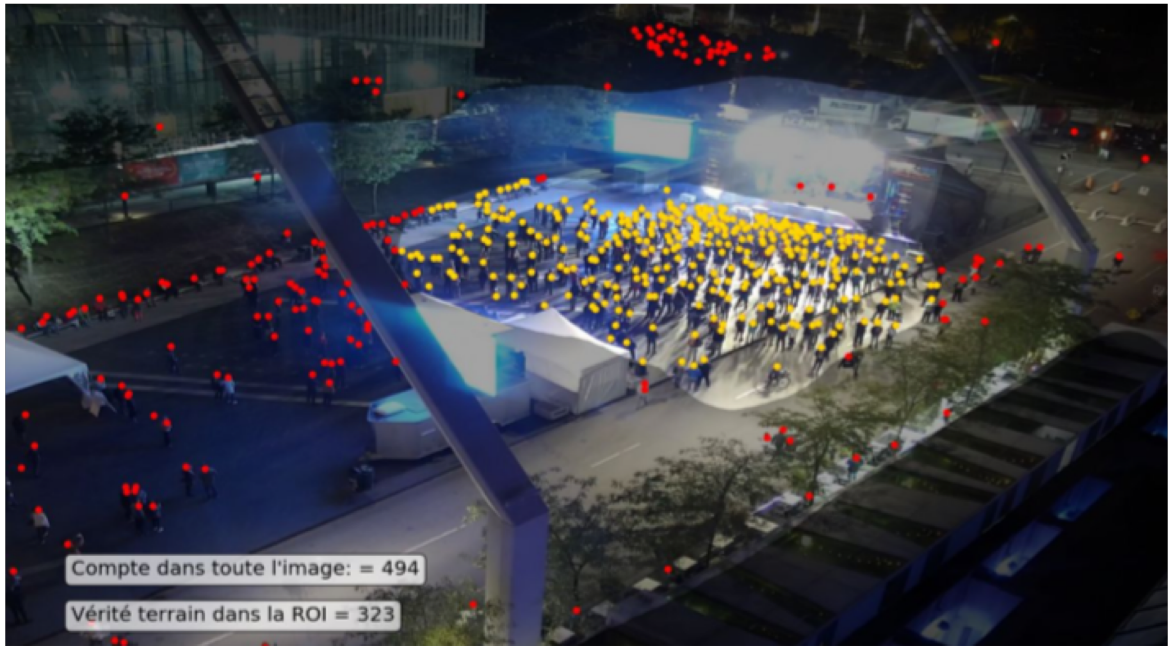
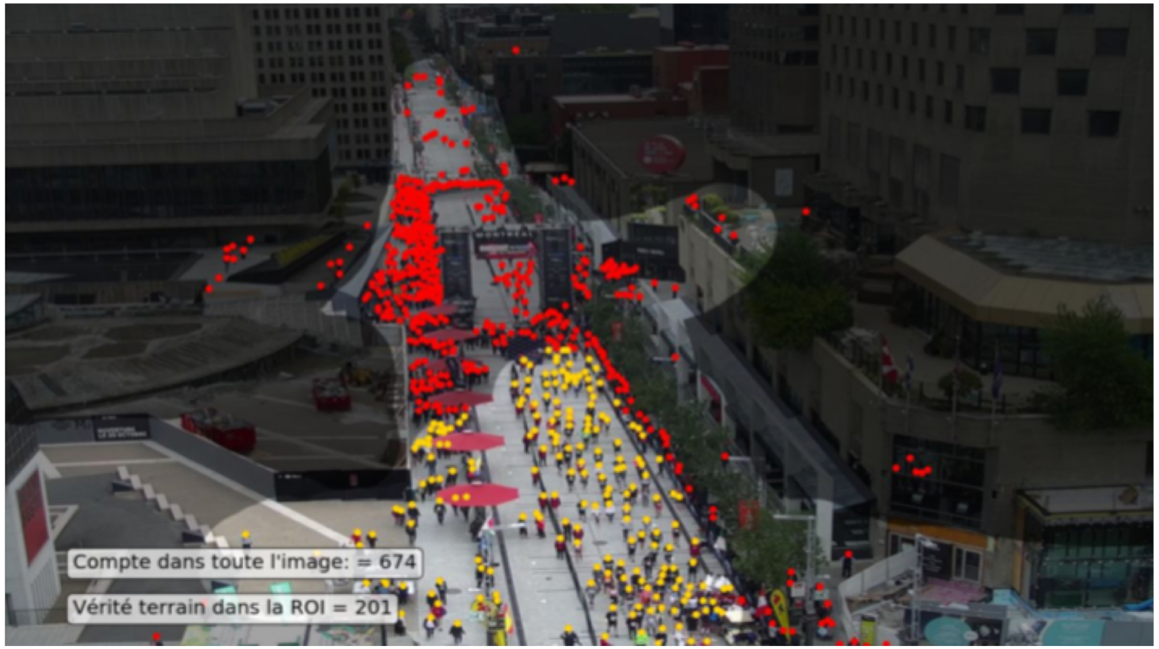


Image credit: https://towardsdatascience.com/

BACKGROUND

An example: count individuals in a crowd


Compte dans toute l'image: = 494
Vérité terrain dans la ROI = 323


Compte dans toute l'image: = 674
Vérité terrain dans la ROI = 201

BACKGROUND

There are very few versatile software or toolkits to annotate datasets.

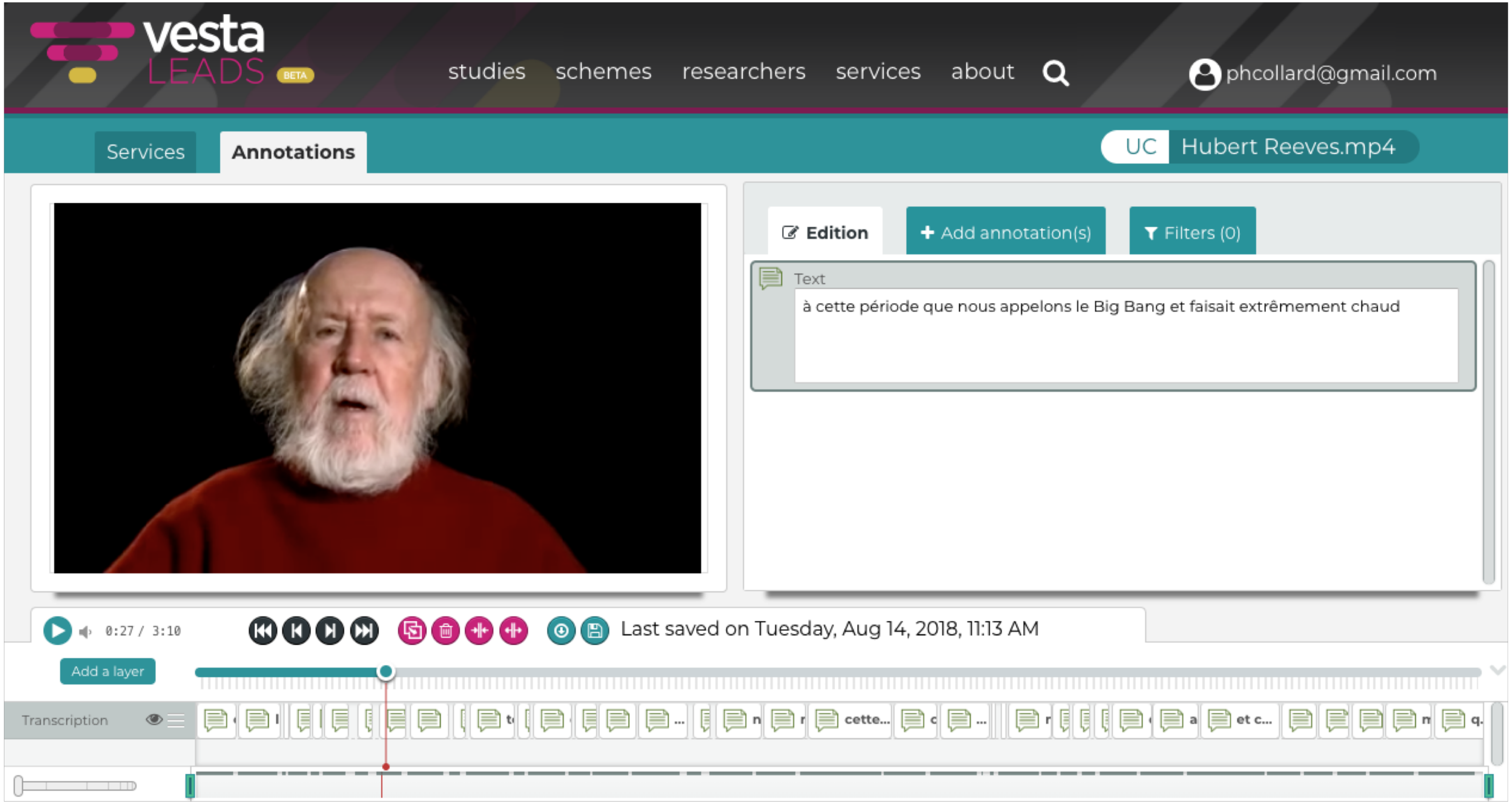
CRIM built two annotation platforms financed by CANARIE Research Software Program.

vesta

PACTE

Those platforms are used by CRIM staff and by clients to annotate datasets and thus building ML training sets for specific projects.

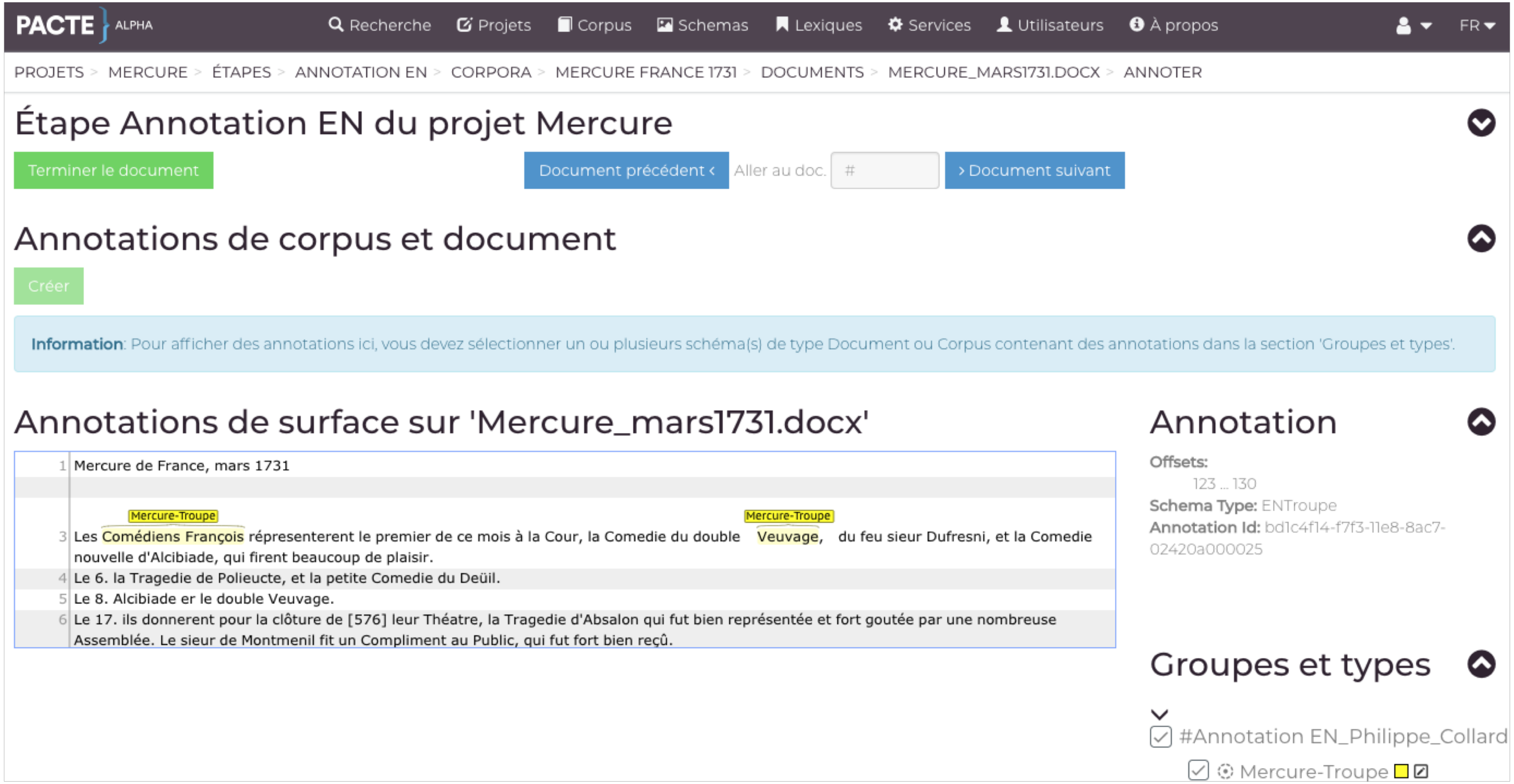
BACKGROUND



à cette période que nous appelons le Big Bang et faisait extrêmement chaud

Last saved on Tuesday, Aug 14, 2018, 11:13 AM

BACKGROUND



Étape Annotation EN du projet Mercure

Annotations de corpus et document

Information: Pour afficher des annotations ici, vous devez sélectionner un ou plusieurs schéma(s) de type Document ou Corpus contenant des annotations dans la section 'Groupes et types'.

Annotations de surface sur 'Mercure_mars1731.docx'

Mercure de France, mars 1731

Les Comédiens François répresenterent le premier de ce mois à la Cour, la Comedie du double Veuvage, du feu sieur Dufresni, et la Comedie nouvelle d'Alcibiade, qui firent beaucoup de plaisir.

Le 6. la Tragedie de Polieucte, et la petite Comedie du Deüil.

Le 8. Alcibiade er le double Veuvage.

Le 17. ils donnerent pour la clôture de [576] leur Théatre, la Tragedie d'Absalon qui fut bien représentée et fort goutée par une nombreuse Assemblée. Le sieur de Montmenil fit un Compliment au Public, qui fut fort bien reçû.

Annotation

Offsets:
123 ... 130
Schema Type: ENTroupe
Annotation Id: bd1c4f14-f7f3-11e8-8ac7-02420a000025

Groupes et types

☑ #Annotation EN_Philippe_Collard

☑ Mercure-Troupe 🟨

BACKGROUND

However...

We do NOT have tools or processes to make our datasets findable, accessible, interoperable, reusable.

We know of research teams who would benefit from an integrated live and static annotation storage engine compliant with FAIR for their work.

FRACS

CANARIE RDM Call 1

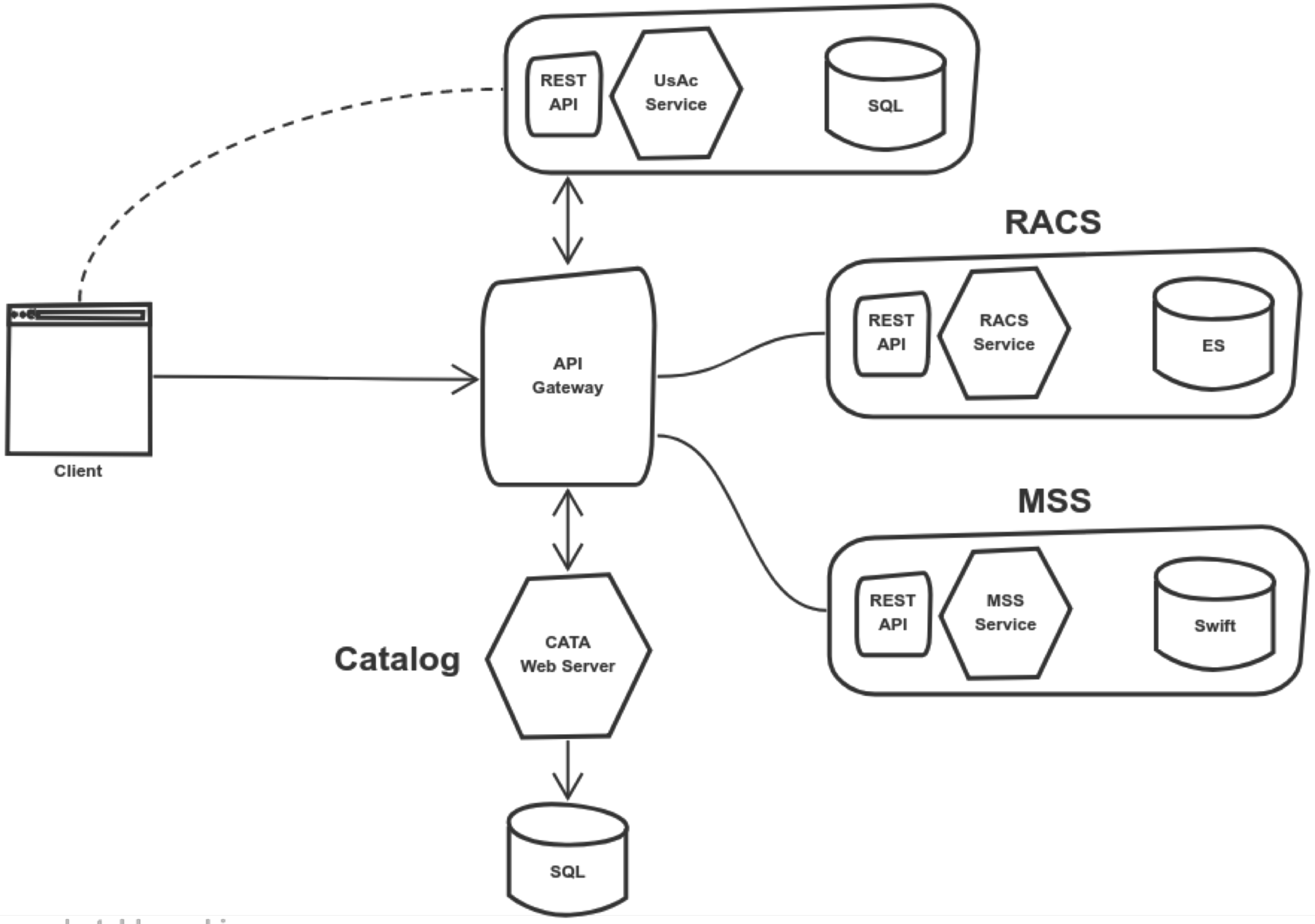
We already had a few components from our annotation platforms

- MSS: Multimedia Storage Service
- RACS: Repository for Annotations, Corpora and Schemas
- UsAc: Users and rights management

“Only” a few blocks missing

- Ability to generate dataset distributions from our live storage engine
- Make our datasets FAIR compliant
- Publish and share them to the world
- Catalog of datasets on the web
- Provision for multi-layered metadata

FRACS



www.sketchboard.io

FAIR APPLIED

FINDABLE

- ❑ **F1 DOIs for distributions**
- ❑ **F2 Multi-layered metadata**
    - ❑ Repository
        - ❑ Catalog
            - ❑ Dataset
                - ❑ Distribution
- ❑ **F3 Unique persistant ID for datasets and metadata**
- ❑ **F4 Public metadata published on the web and indexable**

FAIR APPLIED

ACCESSIBLE

- ❑ A1 REST and GraphQL APIs for metadata and eventually datasets
- ❑ A2 Metadata always accessible even for archived datasets

INTEROPERABLE I1 & I2

- ❑ JSON and XML formats for metadata
- ❑ Visible from REST or GraphQL APIs
- ❑ Vocabulary from common ontologies: DC, Schema.org

REUSABLE

- ❑ R1 Licence and terms of use for every distribution

WHO IS IT FOR

- Research teams in need of a fast and scalable annotation storage engine, either from a simple process or a platform
- Research teams willing to make snapshots or distributions of their live dataset and publish them
- Research teams owning other forms of datasets and inclined to publish them
- Harvesters, search engines
- Researchers looking for datasets
- Researchers publishing papers linking to their datasets
- Developers wanting to re-use one or several components of FRACS
- Anybody sent by CANARIE...

Philippe Collard, M. Sc., PSM

Product Manager - Annotation Platforms

philippe.collard@crim.ca

www.crim.ca

CRIM

101 - 405, avenue Ogilvy, Montréal (Québec) H3N 1M3 514 840-1234 / 1 877 840-2746