RDF datahub for precision medicine

IRISA Rennes : Dyliss + Empenn
CHU Rennes : UF bioinformatique
INSERM U1236 MicMac
Institut du Thorax Nantes
LS2N Nantes : GDD

Data integration is critical for precision medicine

- Patient data are heterogeneous (intrinsically + acquisition modalities)
- Reconciling diseases complexity with patient-specific data ⇒ integrated approach
- Semantic Web technologies = relevant framework for addressing
  - Interoperability
  - Scalability
  - Federation of multiple datasets

IT Challenge: reconcile (1) volume and complexity of data, (2) rich queries, (3) capability to query multiple datasets, (4) acceptable response time

RDF datahub

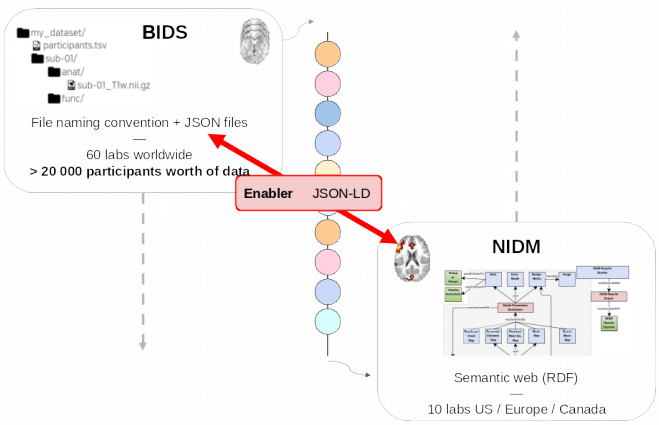
How to query efficiently several complementary datasets?

Compare and combine centralized VS distributed approaches

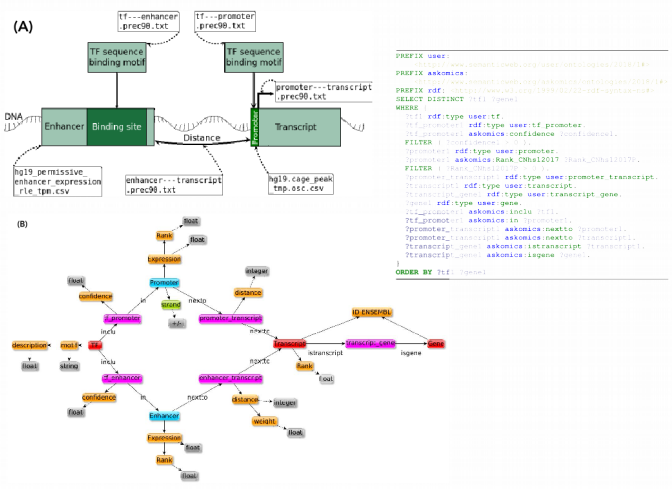
- Identify relevant query scenarios
  - Neuroimaging
  - Whole genome sequencing
- Improve performances of SPARQL queries on centralized RDF datahub
- Improve performances of SPARQL federated queries over multiple datasets

Relevant query scenarios

Neuroimaging: Expose metadata as RDF (NIDM) to facilitate cross-domain queries

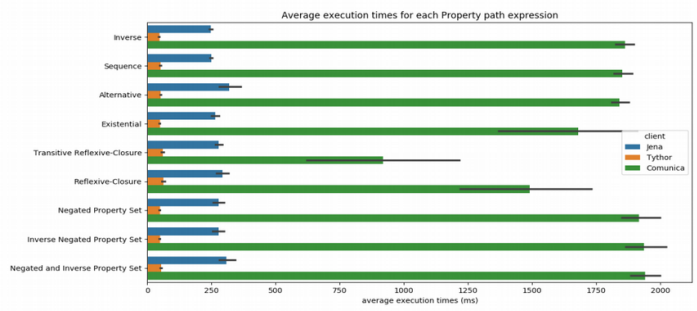


Whole genome sequencing



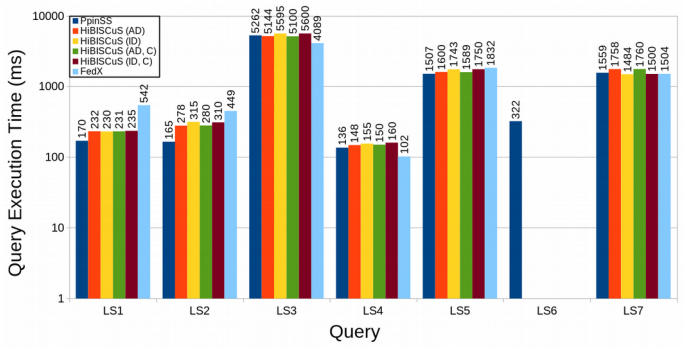
Queries on centralized datahub

- Timeouts often due to property paths and Kleene ops
- Breakthrough:
  - Decompose query on the client side
  - Send a succession of simpler queries
- Gain > 1 order of magnitude
- **Available**: http://sage.univ-nantes.fr/



Federated queries

- SPARQL engine
  - manages the query decomposition into subqueries
  - sends the subqueries to the right endpoints
  - computes the Join and Union of the results
- Breakthrough: a more detailed index supports more efficient query processing



Conclusion

- Semantic Web technologies = relevant framework for representing and integrating datasets
- Improved query performances at the endpoints level
- Improved federated query performances

Perspectives

- Disseminate RDF vocabularies (transitioning from BIDS to NIDM)
- Combine endpoint-specific and federated query improvements
- Disseminate outside of the Life Science community