PEPR D'UNE STRATÉGIE NATIONALE
PROJET CIBLÉ
2022

PEPR SANTÉ NUMÉRIQUE

DOCUMENT PRÉSENTATION PROJET

ShareFAIR

Acronyme	**ShareFAIR**		
Titre du projet en français	Partager des protocoles fiables pour transformer des jeux de données en gold standards : application aux pathologies neuro-vasculaires		
Titre du projet en anglais	Sharing reliable protocols to transform datasets into gold standards: Application to Neuro-Vascular Pathologies		
Mots-clefs	FAIR, workflows, standards, provenance, partage et réutilisation de protocoles, annotation automatique de jeux de données		
Établissement porteur	**Université Paris-Saclay**		
Responsable du projet	**Prénom, Nom, Qualité**		
	Sarah Cohen-Boulakia, Professeure		
	Courriel	**Téléphone**	
	sarah.cohen-boulakia@universite-paris-saclay.fr	**01 69 15 32 16**	
Durée du projet	**48 Mois**		
Aide totale demandée	**1,8 M€**	**Coût complet**	**5 M€**

Liste des établissements du consortium :

Établissements d'enseignement supérieur et de recherche	***Secteur(s) d'activité***
Université Paris-Saclay	***Formation et Recherche***
Université Paris-Dauphine PSL	***Formation et Recherche***
Université Claude Bernard Lyon 1	***Formation et Recherche***
Université Rennes	***Formation et Recherche***

Organismes de recherche	***Secteur(s) d'activité***
Inria	***Recherche***
INSERM	***Recherche***
CEA	***Recherche***
Institut Pasteur	***Fondation de recherche***

Résumé du projet en français (Non Confidentiel – 4000 caractères maximum, espaces inclus)

L'accès à une grande variété de données complémentaires, multi-échelles et massives offre des opportunités sans précédent pour la recherche en santé. Un grand nombre d'analyses peuvent être effectuées sur ces ensembles de données, permettant de faire émerger des avancées scientifiques et des découvertes. La stratégie nationale d'accélération PEPR « santé numérique » ambitionne de stimuler l'innovation en matière de santé numérique, ce qui inclut la conception d'approches innovantes d'analyse des données de santé.

Ces analyses sont complexes, elles reposent sur divers outils qui doivent être paramétrés et chaînés les uns aux autres. Il existe désormais des preuves irréfutables que de nombreuses découvertes scientifiques ne résisteront pas à l'épreuve du temps : améliorer la reproductibilité des résultats obtenus par des approches numériques est d'une importance capitale, en particulier en santé.

Le partage des données de santé est aussi souvent entravé par les impératifs de protection des données personnelles et se heurte à des contraintes techniques (sécurité, volume). Ces contraintes peuvent cependant être limitées lorsque les protocoles et les workflows qui implémentent ces analyses sont suffisamment réutilisables pour reproduire les analyses in situ.

De plus, lorsqu'ils sont conçus pour être réutilisables, les protocoles et workflows fournissent les traces de provenance des données analysées, décrivant comment les résultats ont été obtenus, à partir des données et augmentent ainsi la confiance des scientifiques dans les résultats produits.

Des solutions innovantes pour l'annotation des données biomédicales et cliniques et pour l'extraction de la provenance sont à concevoir. Les protocoles et les workflows qui utilisent et génèrent de grands ensembles de données hétérogènes devraient être élevés au rang d'objets de première classe et la relation duale inhérente entre données et protocoles/workflows devrait être mieux exploitée.

Les défis incluent donc la normalisation et l'annotation des ensembles de données et des protocoles et workflows, l'extraction des protocoles et workflows à partir de données textuelles, cliniques et biomédicales, et leur synthèse en protocoles et workflows interopérables, partageables et réutilisables.

L'originalité de ShareFAIR réside dans le fait d'aborder à la fois la fiabilité des jeux de données et celle des protocoles d'analyse et workflows et d'exploiter la double relation entre les jeux de données et les protocoles. Plus précisément, ShareFAIR fournira

(i) des standards communs et de qualité pour annoter les données, les protocoles, les workflows, et pour fournir une provenance de qualité retraçant l'origine des données,

(ii) un cadre interopérable pour le partage, l'annotation, la réutilisation de protocoles et workflows fiables (FAIR),

(iii) des approches pour extraire des protocoles et workflows à partir de données textuelles afin d'enrichir l'ensemble des protocoles et de mieux documenter la provenance des ensembles de données, et des approches pour apprendre ou extraire des protocoles à partir d'ensembles de données biomédicales et cliniques.

Les preuves de concept et les percées réalisées grâce à ShareFAIR seront appliquées à des cas d'utilisation réels liés aux pathologies neuro-vasculaires avec des ensembles de données multi-échelles (génomique, imagerie neuro-vasculaire et clinique) et des protocoles et workflows d'analyse complexes.

ShareFAIR facilitera la réanalyse des ensembles de données biomédicales tout au long du cycle de vie des projets scientifiques et participera de manière proactive aux efforts à grande échelle vers une science plus reproductible et cumulative. Au niveau de la science des données, ShareFAIR fournira un cadre unique pour la recherche sur l'interopérabilité liée à FAIR. L'objectif et la méthodologie adoptés dans ShareFAIR s'alignent sur les principales infrastructures de recherche européennes telles qu'ELIXIR et EOSC-Life.

Résumé du projet en anglais (Non Confidentiel – 4000 caractères maximum, espaces inclus)

Access to a wide variety of complementary, **multi-scale and massive data collections** offers unprecedented **opportunities for healthcare research**. A large number of analyses can be performed on these datasets, for scientific advances and discoveries to emerge. The national 'Digital Health' Acceleration Strategy ambitions to boost digital health innovation which includes designing innovative health data analysis approaches.

Importantly, such data analyses are complex, they rely on various computational tools that have to be parametrized and chained together. There is now compelling evidence that many scientific discoveries will not stand the test of time: **increasing the reproducibility of computed results** is of paramount importance, especially in the healthcare domain.

Sharing of health data is often hampered by personal data protection requirements and comes up against technical constraints (security, volume). These constraints can however be limited when the protocols and the workflows implementing analyses are sufficiently **reusable to reproduce** analyses in situ.

Additionally, when designed to be reusable, protocols and their implementations - workflows - provide the provenance traces of the analyzed data, describing how data results have been obtained and thus increasing **scientists' confidence in the results produced**.

This calls for **innovative solutions** for the **annotation of biomedical and clinical datasets and extraction of provenance**. Protocols and their implementation as workflows using and generating datasets should be elevated to first-class objects and the inherent dual **relationship between datasets and protocols/workflows** should be better exploited.

Challenges thus include **standardization and annotation for datasets and protocols, extracting protocols and workflows from text and other datasets,** and **synthesizing them into interoperable, yet shareable protocols.**

The originality of ShareFAIR lies in tackling both the reliability of datasets and analysis protocols and in harnessing the dual relationship between datasets and protocols. Specifically, ShareFAIR will provide

(i) **standards to uniformly represent datasets,** ontologies/common vocabularies to annotate datasets and protocols/workflows, and provenance to trace the origin of datasets,

(ii) an **interoperable framework** for the **design, annotation and reuse** of reliable and shareable protocols,

(iii) approaches to **extract protocols** from textual data to enrich the set of protocols and workflows and better document the provenance of datasets, and approaches to learn protocols from biomedical and clinical datasets.

The proofs of concept and breakthroughs reached through ShareFAIR will be applied to real-life use cases related to **neuro-vascular pathologies** with **multi-scale** (genomic, neuro-vascular imaging and clinical) **datasets and complex analysis protocols** and **workflows**.

ShareFAIR will facilitate biomedical datasets re-analysis throughout scientific project lifecycles, and proactively participate in large-scale efforts towards more reproducible and cumulative science. At the data science level, ShareFAIR will provide a unique framework for FAIR-related interoperability research. The objective and methodology adopted in ShareFAIR aligns with prominent European research infrastructures such as ELIXIR and EOSC-Life.

Table des matières

1. Context, objectives and previous achievements

1.1. Context, objectives and innovative features of the project

In the last ten years, we have assisted a deluge of data generated in the field of biomedicine and healthcare. This includes data from a variety of sources, such as electronic health records, genomic data, clinical trial data, and data from wearable devices and sensors. The increasing availability of this data offers unprecedented **opportunities for healthcare research,** it has the potential to revolutionize the way we understand and treat diseases, as it allows researchers to identify trends and patterns that may not have been apparent with smaller data sets. However, it also presents a number of challenges first in terms of data management, for considering the multi-scale, massive and highly heterogeneous aspects of such datasets, and even most importantly the development of new methods for correctly understanding and interpreting the data. Facing numerical challenges posed by biomedical datasets is the main objective of the PEPR Digital Health. More precisely, ShareFAIR belongs to the Program 2 of the PEPR “Tackling the challenges of the uses of multi-scale personalized health data” and it focuses on the axis 5 of this program to help understand and interpret correctly datasets obtained. Here, a major challenge lies in **accompanying both raw and analyzed datasets with their provenance**: *where* they come from, *how* they have been produced, *from which data acquisition system* or *from which input datasets using which analysis protocol*... Provenance is at the center of the FAIR (Findable, Accessible, Interoperable, Reuseable) data principles that become mandatory in Data Management Plans. The ability to **reuse analysis protocols** is crucial for comparing biomedical results, adapting and repurposing protocols to new problems and designing new protocols. However, protocol reuse remains challenging due to the **diversity of data analysis protocols** which may be specified with a variety of executable scripts and sometimes only described textually.

This calls for **innovative solutions** for the **annotation of biomedical and clinical datasets and extraction of provenance**. Protocols using and generating datasets should be elevated to first-class objects and the inherent dual **relationship between datasets and protocols** should be better exploited. Challenges thus include **standardization and annotation for datasets and protocols, extracting protocols from text and datasets,** and **synthesizing them into interoperable, yet shareable protocols.**

The originality of ShareFAIR lies in tackling both the reliability of datasets and analysis protocols and in harnessing the dual relationship between datasets and protocols. Specifically, ShareFAIR will provide (i) standards to uniformly annotate datasets and protocols with ontologies/common vocabularies and provenance to trace their origin, (ii) an interoperable framework to index, design and annotate reliable and shareable analysis protocols, (iii) approaches to extract new protocols, based on the literature, learned from biomedical and clinical datasets, and from international data challenges in neuroimaging.

State-of-the-art

ShareFAIR builds on a number of standards, technologies, and proposals in the area of scientific workflows, data provenance, querying, data and text mining. In this section, we present this landscape, starting with the FAIR initiative, and existing standards, to then move on to existings proposals in scientific workflows, provenance, querying, data and text mining, which together will form the backbone of the ecosystem that will be built under the ShareFAIR umbrella.

FAIR. The 'FAIR Guiding Principles for scientific data management and stewardship' provide guidelines to improve the Findability, Accessibility, Interoperability, and Reuse of digital assets [WDA+16]. The principles emphasize the ability of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention. They mostly rely on associating metadata with datasets to provide a fine description of its contents and origin. ShareFAIR aims to make the most of the FAIR description of the input datasets, protocols and workflows to automatically provide FAIR datasets produced as output.

Analysis Protocols and their implementation. Healthcare institutions make use of different kinds of protocols, two of them are of particular interest to ShareFAIR. First, *care protocols* consist in recommendations of sets of intended exams to be followed by patients to optimize clinical care [WB13,EST+22]. Second, *data analysis protocols* help the management and analysis of large data sets produced in hospitals. Protocols denote here the high level description of the main steps of a clinical pathway or a data analysis process. Concrete implementation of protocols are called *workflows*: the precise process to coordinate multiple modules and their dependencies. Modules can be of very various kinds: e.g., a code to be run, a service to invoke or a command line tool. At run time, the computational workflow is executed: it is fed with input data and all the possible module's parameters are set, then each module executes and produces intermediate and final data. The rise in the use of computational workflows has been accompanied by a range of diverse systems by which they can be implemented. In ShareFAIR we will focus on Workflow Management Systems (WfMS) that provide a feature-rich infrastructure for the definition, set-up, execution and monitoring of a workflow. Pioneer systems include Taverna [WHF+13] or Kepler [LAB+06] but they are no longer maintained. We will focus in ShareFAIR on the next generation of systems Galaxy [GNT10] and more strongly Nextflow [DCF+17] and Snakemake [KR+12].

Standards for representing and annotating protocols, workflows and their components. The specification of interoperable protocols and workflows has attracted the attention of a number of standardization bodies and scientific communities, resulting in a number of specifications. They describe workflows, their constituent steps, data dependencies and, to a lesser extent, control dependencies between the steps. WFDesc [BZG+15], P-Plan [GG12], along with the CWL standard [C+22, A+16], fall into this category. The difference between such proposals lies in the degree of specificity with which they capture workflow description. CWL seeks to capture executable workflows whereas Wf-desc is intended for abstract (non-executable) workflows without getting into the execution details. Formats have also been proposed to represent semi-formally or formally clinical protocols [Pel13], either based on OWL, the Arden syntax or reusing the OMOP common data model [BCB+21].

The community has also produced standards for capturing provenance traces, that is, the traces generated by workflow execution. wfprov [BZG+15] and provOne[CLM+15] are specializations of the Prov-O recommendation of the W3C for provenance traces [MBC13]. RO-Crate [SR+22] and ResearchObjects [BZG+15], act as containers carrying information about the workflow specification, its provenance traces, other types of metadata and auxiliary resources deemed relevant to ensure reproducibility. The above standards are augmented with domain-specific ontologies and vocabularies, to characterize the steps (modules) of the workflows such as the EDAM [IKJ+13] ontology and the datasets used or generated by these workflows and protocols and which are held in the provenance traces of the workflows (e.g., the HPO [GMC+21], GENO[1], SEPIO[2] and MONDO[3] ontologies). As several mature workflow languages are used in bioinformatics, querying workflow repositories require abstraction from these languages, thus leading to the development of workflow interoperability languages. In the business process community, this leads to query languages in the form of abstract processes used as queries, such as in BP-QL [BEK+08].
There is a need to **provide annotation guidelines** for workflow authors and curators in how to use such standards and/or to guide them in which standards to favor depending on their needs. Additionally, there is a need to **cross-link and possibly enrich selected standards** for datasets, protocols, workflows and provenance traces to remove potential gaps, improve coverage and quality, and ensure consistency and sustainability while maximizing the reuse of existing efforts.

Provenance. Workflow provenance refers to the documentation and tracing of the inputs, processes, and outputs involved in the execution of a workflow. It can be used to support a range of functionalities, including debugging workflow specifications, interpreting workflow results, inspecting the quality of the data used and generated by workflow modules, and assessing the reproducibility of the workflow results [DF08,SVR+19]. A number of tools have been developed over the years for capturing workflow provenance, either as a component of a workflow system (e.g., Taverna [WHF+13] or Galaxy [GNT10]), or as a library or toolkit that can be used independently of the workflow system, (e.g., noWorkflow [MBC+14,CMS+20]). The size of provenance captured as a result of workflow execution can be orders of magnitude larger that the input data used to feed workflow executions. Therefore, i) storing provenance information becomes an issue. Furthermore, ii) the sheer size of captured provenance quickly overwhelms users who may be quickly dissuaded from using provenance.
The above calls for **new provenance storage techniques**. The challenge here lies in the ability to store a complete account of the provenance of the data used and generated by the workflow execution without requiring a large storage footprint. In addition, we need to assist users in making sense of potentially overwhelming provenance datasets. In particular, users need to be provided with facts or correlations extracted from provenance traces, which assist them in exploring and making use of provenance information in an straightforward manner. This task can benefit from and inform the annotation of provenance which is the focus of the next section.

Annotation. The dataset harvested through provenance collection can be valuable for a number of tasks that go beyond interpreting the results. For them to be useful, they must be accompanied by

[1] https://www.ebi.ac.uk/ols/ontologies/geno
[2] https://www.ebi.ac.uk/ols/ontologies/sepio
[3] https://www.ebi.ac.uk/ols/ontologies/mondo

semantic annotations that help users to understand and exploit them. As manual annotation can be both tedious and time consuming, means have been investigated for semi-automating this task (such as LabelFlow [ABC+18], and earlier works [MBZ+08]) to exploit the workflow description to propagate annotations among the datasets used and generated by the modules that compose the workflow. Such solutions do have limitations, in that they assume that the modules are annotated, which is not the case for most modules. Also, module annotation can at best be used to have constraints on the actual annotation, i.e., they often fail to identify the exact concept to be used for the annotation. **Therefore, there is a need for new solutions that exploit other sources of information, in addition to the workflow description, to infer data annotation, which can take different forms** (a concept, a term, a description) and **accompany the data sets collected within the provenance traces.** We plan to follow the same approach as other proposals that combine multiple sources of information, e.g., OpenPREDICT, the Evidence Graph Ontology (EVI) [ANL+21] and FAIRSCAPE [LNA+22].

Workflow recommendation and reuse. Studies examined the sharing, reuse and similarity of scientific workflows [CL11,SCL+12,SBC+14], considering Taverna [WHF+13] and Kepler [LAB+06] systems and myExperiment [DCS09] as the platform for workflow sharing. Since then, the landscape of systems and sharing platforms has considerably changed. Workflow developers have moved to collaborative and distributed control-version platforms such as GitHub, for sharing their workflows. Specialized initiatives, such as nf-core [EPF+20], have been developed to define good practices for workflow code, and to curate and homogenize their implementations, using GitHub as a backend repository. Since no distributed workflow search engine currently exists, there is a crucial need to develop means to discover existing workflows. Searches may allow users to retrieve workflows based on very various features such as the set of bioinformatics tools they call or their similarity to another existing workflow. Another important need is related to the huge set of workflows returned by a search query: **clustering similar workflows** may help users to better understand search results [CL11]. More generally, helping users retrieve workflows and combine (possibly part-of) existing workflows to design new ones is key [CL11]. This implies developing solutions to r**etrieve, query and compare workflows** [SBC+14]. While attempts have been proposed in the previous generation of workflows (e.g., [WSL16, SCK+16]), current challenges lie in considering the new shape of scientific workflows: distributed in GitHub projects (not centralized in myExperiments) and made of assembly of pieces of code (not series of named modules).

Extracting information on protocols or workflows from text. Information about shared datasets, software, and workflows implementing protocols is contained in scientific articles. The extraction of information from articles needs to leverage several aspects of information processing and management: knowledge representation, to formally describe the information units of interest, information retrieval to locate the specific documents encoding the information, targeted information extraction (e.g., entity recognition) to extract the specific information units of interest. These tasks remain challenging for natural language processing as they require identifying weak signals (specific information buried in large text corpus) in a weak supervision setting (information of interest has yet to be formalized as an annotated corpus) [NWL+11]. While the development of models for extracting entities and relations relying on annotated data is an area of research despites the advancements brought by Deep Learning methods [WZA+20], the difficulty to obtain annotated data has raised new interest for few-shot learning [LYF+22] and weakly supervised approaches [MBR+09] when dealing with new domains such as the description of biomedical and clinical protocols. Workflow information extraction will leverage these complementary approaches: workflows collected represent knowledge that can be used to produce labeled data through distant supervision in support of few-shot approaches, especially by extracting patterns useful for prompting methods. This association will be used for extracting both workflows' components (e.g., tools, methods), and the relations between them. This line of research will build upon previous work about distant supervision and prompting for named entity recognition [LYH+20], relation extraction [MGL+21,HZD+22] including from the literature [LWY+22,SNH+20].

Learning protocols from data. Another way to discover new protocols is to automatically learn them from data. Protocol execution in real life requires input data and generates output data, leaving traces of executions that can be used to learn protocols. We will consider the case of diagnostic and therapeutic protocols, in the form of clinical pathways that we propose to learn from Electronic Health Records (EHRs). Several methods can be leveraged to reach this goal including sequence and process mining as unsupervised methods, and process mining to discover patterns within hospital processes [BDJ+17]. Resulting constructs can in turn be (i) compared with state-of-the-art clinical pathways or clinical practice guidelines, and (ii) evaluated with regards to quantifiable outcomes related to health (e.g., number of complications), or healthcare management (e.g., the cost). Reinforcement Learning (RL) methods are natural candidates to learning sequential decisions, Deep Reinforcement Learning (DRL) adopts deep neural networks to learn the optimal policy that is central to RL and facilitate handling cases where the number of states is potentially large [LNF19]. To identify optimal treatment regimens for patients, [LWL+22] proposed a deep Q network-based model for EHR data. Patient records are used to model the state, action and transition probability of the model, used, in turn, to determine the individual optimal dose. Note that classical representation formats for clinical protocols such as [BCB+21] are not provenance-aware and they are interoperable only to a limited extent; we will build upon the formats developed in ShareFAIR to represent and compare extracted clinical pathways.

Extracting protocols and workflows from large shared community datasets (neuroimaging). Protocols and workflows can also be extracted from large datasets that are shared within a community and for which the different tools and approaches used are extensive. Such tools are valuable to practitioners and have brought the capacity to process more data in a shorter amount of time. But – each approach provides its own version of the results – and overall approach multiplicity leads to a very large space of possible results leaving practitioners at a loss to find the right answer to their research question. Until recently, this analytical variability induced by different protocols and pipelines/workflows on the results has typically been ignored, considering it as negligible compared to other sources of variability (induced by participants, test-retest, etc.). In 2020 a landmark paper [BN+20] challenged this status-quo in neuroimaging: 70 teams were given the same dataset and asked to answer the same yes/no research questions. Each team chose a given approach leading to contradictory findings. Modeling properly such approaches as FAIR protocols and workflows following the ShareFAIR framework will make it possible to understand the causes of analytical variability and its practical impact on neuroimaging use-cases.

1.2. Main previous achievements

The group of researchers involved in ShareFAIR have obtained several major results.

We have jointly worked on reproducibility of bioinformatics data analysis using scientific workflows. We organized a series of reprohackathons[4] and designed a **comparative framework of scientific workflows for reproducible research** [CBC+17] to benchmark workflow systems. We have additionally worked on workflow similarity, usage of workflows by end-users and developers [CU11,SCK+16,SCL+12,SBC+14]. We have also been pioneers in considering the question of reproducibility results in neurosciences [BMN19] but also text mining [DNN+21].

Members of ShareFAIR have strong experience on how **using scientific workflow systems in various massive analyses of biological data**: NGPhylogeny.fr (between 100 and 1600 executions a day); booster.pasteur.fr, and COVID-ALIGN (for aligning large number of SARS-CoV-2 sequences). Such tools are highly used and are based on the systems that we will consider in ShareFAIR: Nextflow and Snakemake.

Members of ShareFAIR are strongly involved in ELIXIR (European infrastructure integrating and sustaining bioinformatics resources across Europe) and EOSC-Life (an environment for hosting and processing research data to support EU science) and they have been involved in several major International standardization effort: **provenance standardization** (W3C PROV, PROV-O and PROV-DM), refining PROV traces into bioinformatics data summaries, **Common Workflow Language**, ELIXIR **Bio.tools**, evolution of **EDAM**, **FAIR computational workflows** [GCS+20].

We also have strong experience in managing semantic annotation, notably the LERC modular architecture based on a principled organization in RDF named graphs and their own metadata and references to biomedical ontologies [LCG+22].

Last but not least, members of ShareFAIR have actively worked on clinical data for years, in particular in knowledge extraction from patient records [HFF+22,RCR+21].

2. Detailed project *description*

2.1. Project outline, scientific strategy

Access to a wide variety of complementary, **multi-scale and massive data collections** offers unprecedented **opportunities for healthcare research**. A large number of analyses can be performed on these datasets, for scientific advances and discoveries to emerge. The national 'Digital Health' Acceleration Strategy ambitions to boost digital health innovation which includes designing innovative health data analysis approaches.

Importantly, such data analyses are complex, they rely on various computational tools that have to be parametrized and chained together. There is now compelling evidence that many if not most scientific discoveries will not stand the test of time: **increasing the reproducibility of computed**

[4] https://ifb-elixirfr.github.io/ReproHackathon/index-en.html

results is of paramount importance, especially in the healthcare domain. Sharing of health data is often hampered by personal data protection requirements and comes up against technical constraints (security, volume). These constraints can however be circumvented when the protocols and the workflows implementing analyses are sufficiently **reusable to reproduce analyses in situ**. Additionally, when designed to be reusable, protocols and their implementations - workflows - provide the provenance traces of the analyzed data, describing how data results have been obtained and thus increasing **scientists' confidence in the results produced**.

ShareFAIR aims to provide a complete solution for designing FAIR protocols and workflows that can be executed to produce FAIR thus highly reliable datasets. The originality of ShareFAIR lies in tackling both the reliability of datasets and analysis protocols/workflows and in harnessing the **dual relationship between datasets and protocols/workflows**.

In ShareFAIR, we aim to create and foster a **collaborative network** formed by recruited young researchers as PhD students or engineers supervised by at least two different groups of the consortium with complementary and interdisciplinary expertise. The students will be **recruited in a principal host lab, and funding has been planned for research stays in the associated lab(s)** during the project. Annual workshops will be organized as well as datathon and code review sessions. More details are given in Section 3.3.

2.2. Scientific and technical description of the project

In this section we introduce the ShareFair partners and the three main objectives of ShareFAIR (called objectives 16, 17, 18 in the main PEPR document and renamed A to C) and their decomposition into sub-objectives and information on the lead taken by partners. We then present the results of the project. We finally conclude on possible project extensions and on the relationships between ShareFAIR and other axes of the PEPR.

ShareFAIR Partners

ShareFair has a total of seven partners represented with the name of the laboratory followed by the name of the agencies (universities or research institute) that will manage the funding: **LISN-UPSaclay** (Laboratoire Interdisciplinaire des Sciences du Numérique, Université Paris-Saclay), **Institut Pasteur**, **LIRIS-UCBL** (Laboratoire d'Informatique en Images et Systèmes d'Information, Université Claude Bernard Lyon 1), **LAMSADE-UPDauphine** (Laboratoire d'Analyse et de Modélisation de Systèmes pour l'Aide à la Décision, Université Paris-Dauphine PSL), **IRISA-URennes** (Institut de Recherche en Informatique et Systèmes Aléatoires, Université de Rennes), **Inria** (with two groups, **HeKa**, Health data- and model- driven Knowledge Acquisition and **Empenn**), **LIST-CEA** (Laboratory for Integration of Systems and Technology, CEA) and **ITX-INSERM** (Institut du Thorax, INSERM).

Objectives.

Objective A : Define standards to uniformly annotate protocols in terms of analysis tools used and input/output kinds of datasets.

Objective A-1 (Institut Pasteur with ITX-INSERM and all partners): We aim to **identify the set of standards** to annotate workflows with FAIR references, spanning multiple life science data analysis domains from Genomics to Bioimaging, including the description of the biological objects and questions being investigated (e.g. organism, disease, phenotype). The standards selected for use in ShareFAIR will not be used as is, as they may suffer from quality issues such as overlaps, accuracy, and missing concepts (EDAM, EDAM-BioImaging, MONDO, DUO, etc.). We aim to build **a knowledge base** that can be used for annotation and is made up of concepts from existing standards. This will require **adaptation and improvement of concepts borrowed from existing standards to improve their quality and coverage** to address the identified gaps.

Objective A-2 (Institut Pasteur with all partners): Based on (A-1), we aim to develop a **computational solution** to host, query and share FAIR workflows and protocols.

Objective B (17): Define an interoperable framework for analysis protocols. This involves the design, annotation and sharing of reliable and shareable analysis protocols.

Objective B-1 (LIRIS-UCBL with Institut Pasteur): We aim to design a high-level query language **to retrieve protocols, workflows and their executions** that can be translated into lower level queries on the different concrete repositories of workflow languages and execution traces. Queries should include the possibility for expressing provenance requirements (e.g., find workflow using a given dataset) and constraints on workflow tasks (e.g., find workflows using a given bioinformatics tool).

Objective B-2 (LAMSADE-UPDauphine with Institut Pasteur): Our objective is **to capture provenance following a lightweight process to cope with the increasing size of workflow traces**. This objective can be broken down into two sub-objectives. The first focuses on identifying data structures that can be used to index and store provenance with small memory footprints, and the second aims to develop profiling algorithms that can assist users in exploring and exploiting the captured provenance.

Objective B-3 (IRISA-URennes with ITX-INSERM): We aim at **designing a semi-automated FAIRification method** for datasets that will extend low-level provenance metadata with higher-level domain-specific descriptions inferred from the workflow [GSB20]. These descriptions aim at providing a summary focusing on the ``what'' rather than the ``how'', that will be instrumental to the recommendation and interoperability modules. Leverage domain-specific knowledge associated with biomedical datasets, as well as fine-grained workflow execution provenance traces is necessary so that data analysis results can be more easily understood, explained and shared.

Objective B-4 (LISN-UPSaclay with Institut Pasteur): We aim to help the workflow developers **discover, query and compare the large set of workflows** already available in distributed GitHub/GitLab projects, and search for workflows *similar* to a given workflow. Possibly large sets of workflows obtained as a result of such searches and queries should be presented to the users in an intelligible way, grouping workflows into clusters of *similar* workflows.

Objective C (18): Augment the set of current protocols by extracting new protocols from text, and large datasets.

Objective C-1 (LISN-UPSaclay with LIST-CEA): We aim to develop NLP models for **extracting the description of workflows from scientific articles** for matching these descriptions with the implementation of the corresponding workflows when it is available. The description of a workflow should include entities like Tools, Data and the relations between them. Our objective is to perform the extraction of these entities and relations **while minimizing the amount of manually annotated** data.

Objective C-2 (HeKa-Inria with ITX-INSERM): We aim to **learn protocols from clinical data** collected along healthcare activity in Electronic Health Records (EHRs) to explicitize *(i)* medical decision processes, *(*steps to reach a particular diagnosis or therapeutic choice) and *(ii)* management of particular conditions (steps in the management of a particular condition). Protocols extracted from EHRs provide a view on the real-word clinical practice and may then be compared together or with CPG (clinical practice guidelines) which can be seen as more theoretical protocols in that they provide recommendations, or clinical pathways (CP) to standardize clinical practice.

Objective C-3 (Empenn-Inria with Institut Pasteur): Our objective is to **extract protocols and workflows** from large shared community datasets in the neuroimaging domain. We aim to **provide a library of modular workflows and protocols for the neuroimaging community** indexed in the ShareFAIR platform.

The three ShareFAIR results are the following

- **standards to uniformly annotate** datasets and protocols with ontologies/common vocabularies and provenance to trace their origin,
- an **interoperable framework** to index, design, query and annotate reliable and shareable analysis protocols,
- **approaches to extract new protocols**, based on the literature, learned from biomedical and clinical datasets, and from large shared community datasets in neuroimaging.

Relationship between ShareFAIR and other projects of the PEPR

ShareFAIR will be performed in **strong collaboration with ITX- INSERM who leads axis 3 of the Program 3- Cardiovascular "A 5P medicine program to reduce the global impact of stroke"**. More generally speaking, our approach is designed to be adaptable to several cardiovascular or neurological use cases. At the end of the project we could consider pipelines developed in other axes of the program and evaluate how to make them being FAIR workflows and protocols.

Complementary objectives of ShareFAIR

We have identified **three complementary objectives** to ShareFAIR that could be funded by call for expressions of interest (AMI) or call for projects (AAP).

1-*Developing innovative visualization solutions* would be particularly helpful for the comparison of protocols, workflows and for navigating through provenance traces involving annotated datasets.

2-*Benchmarking standardized and annotated analysis protocols and workflows* would allow comparing protocols and workflows in terms of performance and useability. It will be reached by organizing very large data challenges in which various data scientists will (re)use and (re)propose protocols to achieve data analyses (using challenge solutions such as Kaggle, RAMP or Codalab).

3- ***Considering new use cases***, possibly based on very different kinds of data such as clinical trials which may have particular shapes.

2.3. Planning and milestones

In this Section we provide an overview of the ShareFAIR project and introduce the workpackages and the relationships between tasks.

Global view of ShareFAIR

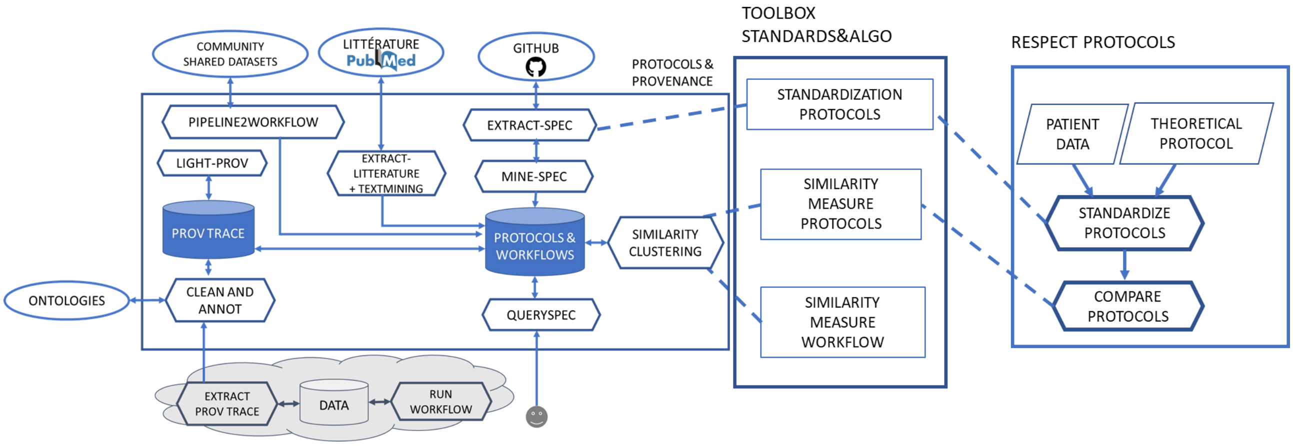

FIGURE 1: Global View of ShareFAIR

Figure 1 provides a global view of ShareFAIR. Workflows and protocols will be extracted from several sources, including GitHub, mined and stored in a knowledge base (WP1.1) to be queried by users (QuerySpec, WP2.1) using similarity techniques for querying and clustering results (SimilarityClustering, WP2.4). Provenance traces will be added to the knowledge base made in WP1.1 to be then reduced and profiled (Light-Prov, WP2.2) and semantically exploited with external ontologies to annotate datasets and turn them into gold standards (Clean and Annot, WP2.3). Additional workflows and protocols will be extracted from articles (extract literature + Text Mining, WP3.1), from large shared community datasets (pipeline2workflow, WP3.3) and from Patient data (respect protocols) where clinical protocols will be standardized and compared. Algorithms for similarity measurement and protocol standardization will be used both for data analysis protocols and clinical protocols (Standards&Algo box).

Organisation of the ShareFAIR project

Figure 2 provides the Gantt diagram of the ShareFAIR project. It summarizes the tasks and underlines the deliverables to be produced. The project is composed of 3 Work Packages: WP1 is in charge of the main implementation part of the project both on reviewing and defining standards and on developing the final version of ShareFAIR platform. WP2 is split into four tasks, each in charge of the design and first implementation of one module of the ShareFAIR platform (langage Module, Provenance Module, Annotation Module, Reuse Module). In each task there is a first subtask to determine the standards to be followed in the module that will be done in collaboration with WP1.1 and a last subtask in which a first version of the module will be implemented. Each module will be consolidated by WP1.2. As for WP3, it will consider additional and more explorative ways of collecting workflows and protocols and will work based on the implementation of standards and modules performed by WP1 and WP2.

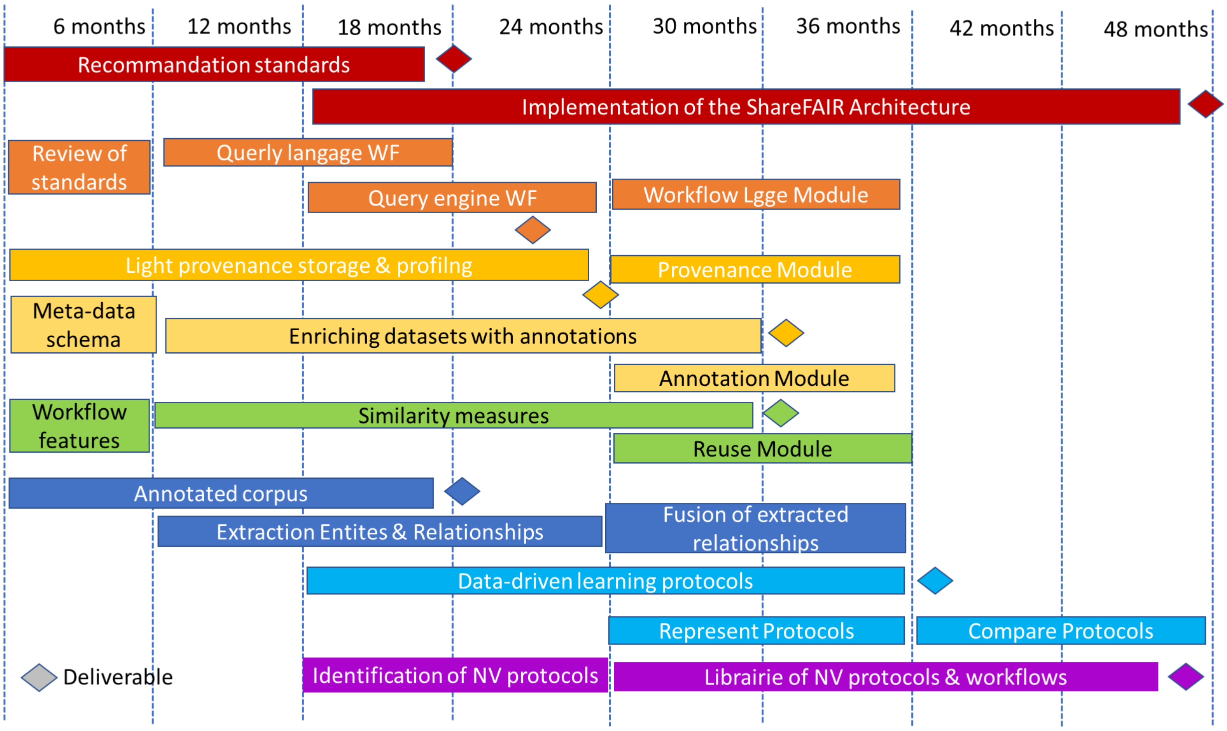

FIGURE 2: GANTT Diagram of ShareFAIR.

In red WP1 (two tasks), in orange WP2.1 (one task divided into 4 subtasks), in dark yellow WP2.2 (one task divided into two subtasks), in light yellow WP2.3 (one task divided into 3 subtasks), in green WP2.4 (one task divided into 3 subtasks), in dark blue WP3.1 (one task divided into 3 subtasks), in light blue WP3.2 (one task divided into 3 subtasks) in purple WP3.3 (one task divided into two subtasks).

Datasets involved

Our target application is the neuro-vascular domain. ITX-INSERM is coordinating the ICAN bio-bank which includes more than 3,400 patients with intracranial aneurysms (ICA). For each patient, a systematic data collection is performed covering clinical observations for around 40 variables (familial history, hypertension, aneurysm size etc.), anatomical imaging (time of flight MRI) and biological samples (DNA). The whole genome sequencing of a clinically relevant subset of 600 patients is under analysis in the context of the WeCAN ANR project. ITX-INSERM manages the data collection with the *Nantes CHU* for clinical data, *France Life Imaging* for neuroimaging data, and the BiRD bioinformatics facility for genomic data. A shareable part of these datasets will be used in the ShareFAIR project in addition to the shareable SnakeMake and Nextflow[5] bioinformatics workflows from ITX-INSERM genomic data analysis. We will additionally work on close datasets that are either public (workflows in github, neuroimaging datasets and pipelines, Pubmed papers) or made available by other partners (hypertension diagnosis protocols and management of intracranial aneurysm from AP-HP and Nantes CHU EHRs).

Work packages and tasks

WP1: Datasets standardization, annotation and provenance. Coordination: H. Ménager (Institut Pasteur). For a set of identified and representative protocols, we will review the state-of-the-art for the annotation of imaging/genomics workflows and protocols, propose coherent and interoperable elements of standardization for annotation of neuro-vascular diseases, **establish a recommended set of annotation standards and a knowledge base** of terms for annotation (**Task 1 - M1-M18**). We will then develop a **computational solution based on such standards** and on modules of WP2 to automatically produce FAIR datasets as output of workflows (**Task 2 - M18-M48**).

Recruitment: Two engineers (named I1 and I2 in the financial document) will be hired in Institut Pasteur (for 23 months each) to respectively provide support in WP1.T1 and WP1.T2.

Deliverables				
D1.1	**M18**	*Guidelines on standards usage*	**Report**	**H. Ménager, Institut Pasteur**
D1.2	**M18**	*Knowledge base*	**Database**	**A. Gaignard, ITX-INSERM**
D1.3	**M48**	*Prototype of the ShareFAIR framework*	**Software**	**F. Lemoine, Institut Pasteur**

WP2: Architecture for Protocols sharing. Coordination: K. Belhajjame (LAMSADE-UPDauphine). This WP aims at specifying the set of modules and the functionalities of the FAIR architecture. It defines the role of each module and the relationship between modules to represent, share and execute FAIR data analysis protocols. It is composed of four tasks.

Task 1 (LIRIS-UCBL with Institut Pasteur). This task consists of five sub-tasks to design the **Workflow language Module.** (i) **Literature review on workflow languages**, specifications in close collaboration with WP1.T1 (M6); (ii) Detailed **identification of workflow and execution representation requirements** (including provenance) together with the assessment of existing workflow and trace stores capabilities (M6); (iii) First proposal of workflow language and execution representation that abstracts at least two workflow execution engines, formalizing semantics of such a language and explaining the mappings between execution engine languages and

[5] https://github.com/lindenb/gazoduc-nf

data/logs/traces and their high level representation (M18: report); (iv) Adaptation/implementation of a basic protocol/trace store containing enough information to answer high-level queries; (v) Implementation of a full query engine in close collaboration with WP1.T2.

Task 2 (LAMSADE-UPDauphine with Institut Pasteur). Task 2 aims at designing the **Provenance Module** using a lightweight solution for ingesting raw provenance traces captured from workflow executions. **Profiling techniques** identifying constraints and correlations underlying provenance traces that can be useful when querying and more generally mining provenance traces will be developed. ***T2.1*** includes exploring techniques for reducing provenance size through **redundancy removal** and **on-demand recomputation** of large datasets, and developing **algorithms for profiling provenance** traces that uncover fine-grained correlations and dependencies between data items within the provenance traces. ***T2.2*** will implement and validate the approach.

Task 3 (IRISA-URennes with ITX-INSERM). Task 3 focuses on designing the **Annotation module** based on the scalable dataset FAIRification method. Inferring more abstract metadata is required for achieving explainable analysis of workflow runs while overcoming their variability. It will integrate low-level metadata describing data, tools and workflows and extend them with higher-level descriptions. T3.1 will **design a multi-view metadata schema** reusing and possibly extending reference ontologies for integrating biomedical data and their processing (specification and execution of tools and workflows). T3.2 will design knowledge-based multi-modal reasoning methods for **enriching the workflow runs with higher level summaries** based on the integration of the data, tools and workflow annotations. T3.3 will assess **how "human oriented" and "machine-actionable" data analysis reports can serve explainability and trustworthiness** issues life scientists are facing.

Task 4 (LISN-UPSaclay with Pasteur). Task 4 provides means for workflow search, reuse and comparison. (i) **Main workflows features will be extracted** (workflow name and authors, workflow graph structure, bioinformatics tools,…) using GitHub metadata and mining information about processors from the workflow files. (ii) **Similarity measures will be identified and new will be designed** to compare workflows and workflow components (processors, tools…). Both syntactic (graph structure) and semantic aspects (e.g., annotations on input and output types) will be considered. We will (a) constitute a gold standard of similarity ratings by experts of the project, (b) identify existing approaches that fit with the next generation of workflows and (c) compare the results obtained by existing approaches with expert ratings. (iii) **Methods to query and recommend workflows** and their components will be designed. Similarity measures will help retrieve workflows similar to a given input workflow and cluster large sets of workflow answers.

Recruitment: Four PhD students (named Doctorant1 to Doctortant4 in the financial document) will be hired in host institutions of the task leaders (UCBL, UPDauphine, URennes, UPSaclay). PhD students will be all co-supervised and provide support respectively to WP2.T1 to WP1.T4.

Deliverables				
D2.1	**M18**	*Report on high level WF and query languages and mapping to low level concrete languages.*	**Report**	**E. Coquery, LIRIS-UCBL**

Deliverables				
D2.2	**M20**	*Specification of the data model and associated indexing structures for ingesting raw-provenance traces, together with algorithms for provenance profiling.*	**Report**	**K. Belhajjame, LAMSADE-UPDauphine**
D2.3	**M24**	*Specification of the ShareFAIR architecture: description of all modules and their relationships*	**Report**	**K. Belhajjame, LAMSADE-UPDauphine**
D2.4	**M30**	*Inference methods for generating annotations summarizing workflow runs.*	**Report**	**O. Dameron, IRISA-URennes**
D2.5	**M30**	*List of similarity measures with their implementation status and preliminary results on workflows.*	**Report**	**S. Cohen-Boulakia, LISN-UPSaclay**

WP3: Extracting and discovering protocols. Coordination: A. Coulet (HeKa-Inria). This WP aims to augment the set of protocols and workflows by considering (i) workflows in publications, (ii) protocols extracted from datasets and (iii) workflows from large shared datasets from the neuroimaging community.

Task 1 (LISN-UPSaclay with LIST-CEA) is composed of three subtasks. **(1) Development of a corpus of workflows descriptions from scientific articles and annotation of a subset of the collected corpus (M1-M18).** The objective of this task is both (i) to **develop tools for crawling scientific articles** describing workflows and to use these tools for collecting a corpus of such articles and (ii) to **annotate** a subset of the collected corpus for evaluating methods developed for extracting information about workflows from scientific articles. Annotation will focus on the entities that are part of workflows and the relations between them. **(2) (M6-M24): Extraction of workflows entities and relations from scientific articles** by minimizing the need for manual annotation through the association of information extracted from GitHub and zero or few-shot learning models relying on pretrained language models. The task will explore how to exploit data coming from distant supervision in these models. For the relationships, we will consider the relations between workflow's entities, possibly at the intra-sentential level. **(3) (M25-M36): fusion of extracted information for building workflow representation.** We aggregate information extracted at the local scale for building a global representation of the workflow described in a scientific article and link it, when it is possible, to a collected workflow.

Task 2 (HeKa-Inria) Learning new protocols from clinical data (M12-M48). This task consists of three steps. **(1) Development of methods for data-driven protocol learning (M12-M36).** We will start with synthetic datasets and the diagnosis of hypertension from EHRs and consider pathways to various sub diagnoses such as primary *vs.* secondary hypertension. We will study the robustness of the pathways extraction to variability and incompleteness of data. Experimentation will then be performed on real-world cardiological data of the AP-HP. Two interpretable approaches will be compared: decision trees, and Q-learning. **(2) Representation of learned protocols with FAIR standards (M24-M36).** We will define a common representation framework of clinical protocols by leveraging the experience of FAIRShare on scientific workflows into a shareable, traceable, understandable format that supports reasoning mechanisms. **(3) Comparison of clinical pathways associated with the same protocol (M36-M48).** Clinical practice is inherently variable, depending both on patient individual history and clinical setting. We will compare such variants with various metrics: Pathways to diagnosis will be compared in terms of number of steps (*i.e.,* tests, questions), time before diagnosis and cost, while pathways for disease management will be compared in terms of observable clinical outcomes (e.g., survival). After hypertension diagnosis, we will focus on neuro-vascular pathologies, by learning and comparing variants in the management of intracranial aneurysms, with the goal of guiding the establishment of future expert recommendations in this domain.

Task3 (Empenn-Inria): Towards FAIR Neuroimaging workflows and protocols (Empenn). This task aims to apply the methods developed in WP1/2 to the field of neuroimaging. The NARPS project [BotvinikNezer2020] where 70 teams were given the same neuroimaging dataset and tasks to answer the same yes/no research questions will be our main testbed. Data and method description of each team in compliance with the COBIDAS guidelines [Nicholsetal2017] are open. We may consider two additional large shared datasets namely (i) *Openneuro,* an open repository that gather datasets from 770 neuroimaging studies (https://openneuro.org/) and (ii) *historical open datasets* widely used in the neuroimaging community as the Alzheimer's Disease Neuroimaging Initiative (ADNI) and its 3,740 publications https://adni.loni.usc.edu/news-publications/publications/**.** We will work on two subtasks. First (M12-M24), we will **extract a set of protocols** from the semi-structured description provided by each of the teams to provide a high-level description of the methodology used. Then (M24-M48), **we will design a set of FAIR workflows and protocols.** This will allow us to model fMRI analysis pipeline variability.

Recruitment: Three PhD students (named Doctorant5, Doctorant6 and Doctortant7 in the financial document) will be hired respectively in task leaders' labs (Heka-Inria, LISN-UPSaclay, Empenn-Inria) to provide support respectively to WP3.T1, WP3.T2 and WP3.T3.

Deliverables				
D3.1	**M18**	*Corpus of workflow descriptions, partly annotated.*	**Report & dataset**	**A. Névéol, UPSaclay, LISN-UPSaclay**
D3.2	**M36**	Method for extracting clinical pathways from EHR data	**Report**	**A. Coulet, Heka-Inria**
D3.3	**M48**	*Set of neuroimaging FAIR workflows and protocols represented in the ShareFAIR framework.*	**Report & dataset**	**C. Maumet, Empenn-Inria**

3. Project organization and management

3.1. Project manager

Sarah Cohen-Boulakia is a full Professor at Université Paris-Saclay. She holds a Ph.D. in Computer Science and a habilitation from the University Paris-Sud. She has been working for twenty years in multi-disciplinary groups involving computer scientists and biologists or physicians of various domains. She spent two-years as a postdoctoral researcher at the University of Pennsylvania, USA and 18 months at the Institute of Computational Biology (IBC) of Montpellier, France. She has long-term collaborations and several short term stays (2 to 4 months) notably at the University of Manchester and Humboldt University zu Berlin, which are major scientific actors in the domain of analysis of scientific datasets using scientific workflows.

In the last years, she has animated several working groups on reproducibility of scientific experiments. She has also been strongly involved in the European Research Infrastructure ELIXIR (https://www.elixir-europe.org/) and on the FAIR initiative (co-organization of a session in the International FAIR Convergence Symposium).

From the National perspective, she heads the CNRS GDR (National research cluster, 1,200 members) MaDICS which is in charge of research animation in the interdisciplinary domain of data science. At local level (Paris-Saclay), she is Vice-Director of DATAIA, the Artificial Intelligence Institute of the University Paris-Saclay where she animates the research on IA for the life sciences in particular by organizing challenges, hackathons and datathons. She is also the head of the Master of Bioinformatics at the University Paris-Saclay (50 to 60 graduates per year).

Dr. Cohen-Boulakia's research expertise include provenance in scientific workflows systems, reproducibility of scientific experiments, integration, querying and ranking in the context of biological and biomedical databases. She is part of the Database and Bioinformatics communities. She has co-supervised 8 PhD students and 16 master students. As of Jan 2023, she has a h-index of 25 with a total of 3100+ citations. She has recently been invited to give keynotes and talks at the national level (EGC 2022 - Extraction et Gestion des connaissances and Journée Reproductibilité de la recherche SIF Société Informatique de France) and international level (Workshop on FAIR Computational Workflows - ECCB 2020; Workshop on Provenance for Transparent Research T7, 2021) on the topic of ShareFAIR. Since 2020, she has additionally been strongly involved in the covid-nma project (https://covid-nma.com/about-us/operating-team.php)– supported by the WHO and Cochrane -- where she leads the data integration group who is in charge of integrating into a datawarehouse the set of all the clinical studies on COVID-19.

3.2. Organization of the partnership

This consortium provides a **unique combination of complementary competences** in computer science: experts provenance in scientific workflows in LAMSADE-UPDauphine/LISN-UPSaclay/ITX-INSERM, databases-data standards experts in LIRIS-UCBL/Institut Pasteur/Empenn-Inria, knowledge representation in IRISA-URennes/LISN-UPSaclay/LAMSADE-UPDauphine, text and data mining in LISN-UPSaclay/LIST-CEA/HeKA-Inria. In the health domain, the consortium gathers **physicians, geneticists and experts in neuro-vascular pathologies** notably in ITX-INSERM, Institut Pasteur, HeKA-Inria and Empenn-Inria. Most importantly, members of this project have a strong experience in multi-disciplinary work at the interface of Computer sciences/Applied mathematics and Health/Biology. **Each partner has also already worked with between one and five other partners of ShareFAIR in the last five years**.

The project relies on access and relationships that partners have on **European Infrastructures** (e.g., ELIXIR, EOSC Life) and **National Research Infrastructures** (e.g., Institut Français de Bioinformatique, France Life Imaging, Plan France Médecine Génomique). More precisely, **ITX-INSERM** partner is also operating the BiRD bioinformatics facility, an HTC infrastructure aimed at accelerating the processing of large scale genomics data. BiRD consists in ~1000 CPUs, ~4TB of RAM and 1PB of storage and is a member platform of the IFB national bioinformatics infrastructure. In addition, **Institut Pasteur** maintains a highly used instance of the Galaxy workflow system, and a large HPC infrastructure consisting of ~15000 cpus+gpus.

We now focus on partners and members of the project (focusing on thesis/engineer supervisors).

LISN-UPSaclay (Part1-Coord) brings strong expertise in biological data management (Data science department, **Bioinfo group**) and text mining on biomedical literature (Sciences and Language Technologies department, **ILES group**). **Aurélie Névéol** is expert in clinical and biomedical Natural Language Processing, addressing both methods and applications of biomedical text analysis, ranging from explorations of representation models and their cross-domain adaptability, to the integration of representation frameworks to extract new medical knowledge from clinical text.

Institut Pasteur (Part2) has a strong expertise in databases, data standards, and development and execution of biological workflows. In particular, **Hervé Ménager, leader of WP1,** is the co-leader of the Bioinformatics and Biostatistics Hub at Institut Pasteur, dedicated to bring support in computational biology to Institut Pasteur's research Units and platforms. He was one of the designers and developers of Mobyle, a bioinformatics workflow framework, one of the contributors of the bio.tools software tools registry, and is nowadays highly involved in the ELIXIR network. **Frédéric Lemoine** has a strong experience in managing and analyzing massive biological data with workflow-based tools such as NGPhylogeny.fr (300+ citations, 100-1,600 executions/day), booster.pasteur.fr (360+ citations), or COVID-ALIGN.

LIRIS-UCBL (Part3) provides strong expertise in database management and in linking databases with languages. **Emmanuel Coquery** has experience in designing large data warehouses for COVID data (Covid-NMA project) and querying very large datasets. He very interestingly combines theoretical and practical skills in databases.

LAMSADE-UPDauphine (Part4) conducts research around the design, the use and the validation of decision. **Khalid Belhajjame, leader of WP2,** is a leading expert in the field of data and knowledge management. He has dedicated his career in these areas has led him to develop innovative solutions that advance open science, FAIR, and reproducibility in modern science, particularly in the life sciences. He has played a critical role in several major International projects, including myGrid, Wf4ever, and DataONE, that have resulted in popular systems and infrastructures for workflow and provenance management. These include the Taverna workflow system, myExperiment, the YesWorkflow system, and Research Objects. Dr. Belhajjame has also contributed to the development of several standards, including the W3C Prov recommendations for provenance and ProvONE for scientific workflows, and made significant contributions to provenance benchmarking through ProvBench.

IRISA-URennes (Part5) provides expertise in biological knowledge-representation in particular in methods based on ontologies for analyzing biomedical data. **Olivier Dameron** is an expert in exploiting symbolic domain knowledge for improving the integration and analysis of large, complex, highly interdependent and often incomplete datasets. He uses Semantic Web technologies (such as RDF, SPARQL and OWL) for integrating distributed data and for combining different kinds of reasoning such as deduction, classification or comparison.

Inria (Part6). HeKa-Inra has a long term experience in knowledge extraction from clinical data, in particular with the development of deep phenotyping tools, i.e., tools that enable retrieving sets of patients with a common profile, out of complex healthcare data. To this matter, the HeKA team is developing a Python library named medkit, dedicated to the extraction of patient features from clinical data. **Adrien Coulet, leader of WP3,** has a successful experience in knowledge discovery from complex biomedical data, in particular by leveraging knowledge representation to guide the extraction process. He was PI of the ANR PractiKPharma project, which focused on knowledge discovery and comparison for pharmacogenomics and PI of the Snowball Inria-Stanford Associate team, which focused on predicting drug response variability.
Empenn-Inria provides expertise in medical imaging, neuroinformatics and population cohorts. In particular, Empenn targets the detection and development of imaging biomarkers for brain diseases and focuses its efforts on translating this research to clinics and clinical neurosciences at large. **Camille Maumet** is an expert in neuroimaging reproducibility. Her research focuses on the variability of analytical pipelines and its impact on our ability to reuse (and use) brain imaging datasets. She is also strongly involved in open science.

ITX-INSERM (Part7) brings expertise in producing and managing massive biological (e.g., whole genome sequencing) and clinical datasets involving several thousands of patients. **Alban Gaignard** co-leads the IFB interoperability working group and represents the french activities in this domain for the European Elixir infrastructure. He is a member of the BiRD bioinformatics facility for genomic data.

LIST-CEA (Part8) brings long term expertise in Natural Language Processing. **Olivier Ferret** is an expert in this domain in particular in learning of semantic and topical knowledge from texts, extraction of semantic relations and Information extraction.

3.3. MANAGEMENT FRAMEWORK, BUDGET, INDICATORS, RISKS

This section provides details on the organization of the project, it then reviews the indicators of success of the ShareFAIR project and discusses the risks associated with the project.

Organization between partners - Budget.

As mentioned previously, ShareFAIR gathers partners of very complementary expertises who have already successfully worked together. ShareFAIR will hire 7 PhD students and 2 engineers, all of them will be co-supervised by both computer scientists (CS) and life scientists (LS). Co-supervision will be performed between different groups with the exception of Inria-HeKA and Inria-Empenn who are mixed teams including both CS and LS experts.
Each team welcoming a PhD or an engineer will be provided with a **3K€ master internship** to start with a 6 months internship of M2, **5K€ for the student to spend time in another partner** (at ITX-INSERM in Nantes, at IRISA-URennes or Empenn-Inria in Rennes, at LIRIS-UCBL in Lyon and at LISN-UPSaclay/LIST-CEA/ LAMSADE-UPDauphine /Institut Pasteur in Paris area) and for the student to travel to at least one International conference during the PhD, **2K€ of hardware** (laptop and some connectors associated) and a total amount of **12K€ for the 4 years of the**

project to cover publication fees and travel expenses/publication fees for the permanents members of the team. More precisely, the 12K€ has been placed in totality in travel expenses to present publications for computer-sciences groups (ILES@LISN-UPSaclay, LAMSADE-UPDauphine, IRISA-URennes, LIRIS-UCBL) while some groups (Bioinfo@LISN-UPSaclay, HeKa-Inria, Empenn-Inria, Institut Pasteur) publishes both in conferences and in Journal (with publications fees) so that the 12K€ have been split into travel expenses (8K€) and publication fees (4K€).

All members of ShareFAIR will meet virtually every two months and physically **twice a year for coding sessions/hackathons** (**18K€** in total, **2,25K€ per datathon** covering catering for 15 persons for 1,5days, travel expenses are covered by partners). One day dedicated (workshop) to all students and young researchers involved in the project will be organized (**12K€** in total, **3K€ per year** covering catering for 30 to 50 persons for 1 day). Such days will be organized in institut Pasteur-Paris (year 1), in Rennes (year 2), in Paris-Center (year 3) and in Nantes (year 4). In year 2 and 4 one public day (**30K€**, 15K€ each day with catering for 80 to 100 participants) will be organized in Saclay to present intermediate and final results obtained in ShareFAIR.
Budget associated with the organization of all events (hackathons, workshops and public days) will be managed by LISN-UPSaclay.

Responsibilities will be shared among the partners.

S. Cohen-Boulakia, coordinator of the project, will be responsible for the organization of both virtual and annual meetings, animation of the project and coordination of reports. As for the WPs, **H. Ménager** will coordinate WP1, **K. Belhajjame** WP2 and **A. Coulet** WP3. **F. Lemoine** will be in charge of ensuring the technical coherence of the implementation choices. **A. Gaignard** will be in charge of the organization of datathons. **E. Coquery** will design and update the website of the project. **O. Dameron** will be in charge of the coordination of the organization of all physical meetings. Recruited engineers will help in the project management.

Indicators of success.

To measure the results of our project we will use three key primary indicators

- The **number of publications involving several partners of the project** will demonstrate how effective the collaboration between partners is.
- The **number of protocols and workflows involved in the ShareFAIR platform** and made available to the community will serve as an indicator to our approach to scale. Quality indicators will be associated with protocols and workflows to characterize them.
- A last indicator consists in **reporting the extensions of already adopted community standards**.

However in the first phase of the project such indicators will progress slowly, thus we will use another set of indicators to measure the dynamic of the project and the alignment of the different partners.

- The **mobility**, both in terms of domain and location, of researchers and especially young researchers as another indicator of effective collaboration.
- The number of **recommendation standards** validated across the different work packages,
- The number of **pairs of modules** of the ShareFAIR platform that are able to deliver joint results (e.g. provenance module, annotation module…)

We undertake to provide such information to ANR upon request.

Risks

The members of our consortium already worked together and obtained several first contributions to rely on. However, we are well aware of potential risks and drawbacks inherent in any research project, and below we comment on those that we have identified.

Recruitment risks: We plan to recruit 7 PhD and 2 engineers. Recruiting many PhD students with the appropriate interdisciplinary background, given the international and national competition for recruiting students and researchers, remains a challenging task. However, partners (especially from universities) are strongly involved in local Master's programs and the recruitment of several good master's students that want to perform an internship and then a PhD is already being processed. A few promising doctoral candidates have indeed been already identified. A risk is then if the PEPR is not ready to be launched by June 2023 but a few months after. Then we may loose the promising National candidates and need to postpone a large part of the recruitment the year after to have excellent candidates that apply for PhD in May. As for the two engineers to be recruited, Institut Pasteur have long term experience in such recruitments. As this project is thought for 4 years, some flexibility on the recruitment dates of engineers will be possible.

Availability of datasets: Access to neuro-vascular datasets may be complex for privacy and security reasons or because all datasets are being produced and are not all yet available. To address this risk, each task in each WP has clearly identified the preliminary datasets close to the target datasets that will serve as a testbed, considering either already available datasets from ITX-INSERM or similar datasets from local partners or public datasets (see *Section 2.3, Datasets involved*).

Work in silos: Given the tasks that will run in parallel and the number of people and partners involved, a risk would be for each team to work in silo (independently of the others) and realize too late the challenges to face to reunite each partner's work. This is the reason why our GANTT guarantees at each phase of the project that deliverables can work together and why we have people in charge of the implementation coordination and organization of code review. This is also why we introduced several indicators to measure cooperation in the project. Moreover our joint experiences in previous projects show good cooperation within and between partners, and we cannot foresee any unresolvable cooperation issues.

Development effort: The prototype to be provided at the end of the project is both complex and ambitious. We have then planned to hire two skilled engineers and hire only PhD students skilled in development, as the part of development during their PhD will be important. Several permanent engineers have also joined the project to help us develop specific implementation tasks in all partners (Cyril Grouin form LISN-UPSaclay, Françoise Conil from LIRIS-UCBL, Jeanne Got from IRISA-URennes, Thomas Cokelaer from Institut Pasteur, Elise Bannier from Empenn-Inria). As already mentioned, three permanent engineers will be strongly involved (Hervé Menager, Fréderic Lemoine and Alban Gaignard) in the project and be in charge of coordinating each (development) task of WP1 and animating the code review (hackathon) sessions.

3.4. Institutional strategy

ShareFAIR is coordinated by University Paris-Saclay and has seven other partners.

ShareFAIR places at the center of the societal challenge “Health and well being” defined as one of the eight prior research domains of the **University Paris-Saclay** which is additionally strongly involved in Open Science and have defined an Open Science Roadmap, promoting FAIR principles.
As an international center for biomedical research, **Institut Pasteur** research domains are pluridisciplinary, from biology to bioinformatics, with several applications to human health. It is highly involved in open science and good practices for reproducible science[6].
University Lyon Claude Bernard 1 is a multidisciplinary university specializing in both fundamental and applied research. It has numerous original developments in cutting-edge fields including medicine.
The **University of Rennes** is fully committed to the open science and scientific integrity movements and has both a collegium of health and a collegium of sciences[7]. **Inria** has listed Digital Health[8] as one of its priority research axes in his strategic plan (*COP 2019-2023, Contrat d'Objectif et de Moyen*) and co-pilot the implementation of the PEPR Digital Health.
University Paris-Dauphine PSL is founded on six disciplines including computer-science and mathematics and it fosters the emergence of high-level research projects, both within and across disciplines[9]. Paris-Dauphine coordinates the IA2-axis 4 project of the program PEPR Digital health.
INSERM is the major French research institute on health, co-leader of the PEPR.
Health and Life Science research is one of the major research areas of **CEA[10].**

4. Expected outcomes of the project

This section will review the impacts and benefits of the ShareFAIR project.

Society. ShareFAIR aims to improve knowledge transfer among geographically-distributed groups of medical professionals by making various clinical and analysis protocols reliable and shareable. ShareFAIR will facilitate transparency of medical decisions with traceable facts to allow stronger treatment acceptability by patients and develop 5-P medicine (personalized, preventive, predictive, participatory, and proof based). This will enable the processing of patient data using these protocols, providing individualized benefits for patients. This will not only benefit patients by increasing the acceptability of treatments, but also enhance the reliability and accuracy of clinical guidelines, ultimately improving the overall quality and cost-effectiveness of healthcare.

[6] https://www.pasteur.fr/en/ceris/library/committing-open-science
[7] https://www.univ-rennes.fr/en/departments-schools-and-institutes
[8] https://www.inria.fr/en/digital-health
[9] https://dauphine.psl.eu/en/research/research-at-dauphine
[10] https://www.cea.fr/english/Pages/research-areas/health-and-life-sciences.aspx

Science. At the life science level, ShareFAIR will facilitate biomedical datasets re-analysis throughout scientific project lifecycles, and proactively participate in large-scale efforts towards more reproducible and cumulative science. At the data science level, ShareFAIR will provide a unique framework for FAIR-related interoperability research. **The objective and methodology adopted in ShareFAIR aligns with prominent European research infrastructures such as ELIXIR and EOSC-Life[11].**

Impact on Research Training. ShareFAIR will result in added-value to the academic institutions at each partner's site in the form of training of master's students and young researchers. In addition, the results of ShareFAIR will be communicated in seminars led by the project consortium, in which the students will be active participants.

Dissemination and exploitation. ShareFAIR will make contributions of different types (standards, datasets, algorithms and tooling). This will give rise to deliverables, published papers, developed software and benchmarks, which will be made available on the project website. Results will be disseminated essentially through scientific publications. The vocation of our tools is to be accessible to the biological/medical and computer science scientific communities and will be made available to all. Moreover, our experience suggests that making software products open is a good way to disseminating and improving the adoption of research results. We plan to develop our software in open source under a permissive license (e.g., LGPL,CeCILL-C) and host it within a public code repository (github, gitlab). As for the intellectual property of the software developed within ShareFAIR, we plan to make an APP (Agence pour la Protection des Programmes) deposit to ensure that each partner retains ownership of the part it has developed.

Through the ShareFAIR consortium, we anticipate the opportunity to **form a network of collaborators not only in France, but also throughout Europe.** We believe that this collaboration will not only enhance the quality of our research and promote the uptake of our results, but also position us as leaders in the field of FAIR data management on a global scale.

References

[A+16] P. Amstutz et al. Common workflow language, v1.0. Figshare https://doi.org/10.6084/m9.figshare.3115156.v2 (2016).
[ABC+18] P. Alper, K. Belhajjame, V. Curcin, C. A. Goble: LabelFlow Framework for Annotating Workflow Provenance. Informatics 51: 11 (2018)
[ANL+21] S. Al Manir, J. Niestroy, M.A. Levinson, et al (2021). Evidence Graphs: Supporting Transparent and FAIR Computation, with Defeasible Reasoning on Data, Methods, and Results. In: Provenance and Annotation of Data and Processes. IPAW 2021. LNCS vol 12839. Springer.
[BCB+21] Boudis, F., Clement, G., Bruandet, et al. 2021. Automated Generation of Individual and Population Clinical Pathways with the OMOP Common Data Model. In Public Health and Informatics (pp. 218-222).
[BEK+08] C. Beeri C, A. Eyal, S Kamenkovich et al. Querying business processes with BP-QL. Information Systems. 33(6):477-507, 2008.
[BMN19] A. Bowring, C. Maumet, and T. E. Nichols, "Exploring the impact of analysis software on task fMRI results," Hum. Brain Mapp., vol. 40, no. 11, Art. no. 11, Aug. 2019, doi: 10.1002/hbm.24603.

[11] https://www.eosc-life.eu/

[BN+20] R. Botvinik-Nezer et al. “Variability in the analysis of a single neuroimaging dataset by many teams”. en. In : Nature 582.7810 (juin 2020), p. 84-88. issn : 1476-4687.

[BS+13] S Bechhofer, et al. Why linked data is not enough for scientists, Future Generation Computer Systems, Vol 29(2) 2013, pp 599-611, ISSN 0167-739X

[BZG+15] K. Belhajjame, J. Zhao, D. Garijo, M. Gamble, K. M. Hettne, R. Palma, E. Mina, Ó. Corcho, J. M. Gómez-Pérez, S. Bechhofer, G. Klyne, C. A. Goble: Using a suite of ontologies for preserving workflow-centric research objects. J. Web Semant. 32: 16-42 (2015)

[C+22] M Crusoe et al. Methods included: standardizing computational reuse and portability with the Common Workflow Language. Commun. ACM 65, 6 (June 2022), 54–63. https://doi.org/10.1145/3486897

[CBC+17] S Cohen-Boulakia, K Belhajjame, O Collin et al: Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. FGCS. 75: 284-298 (2017)

[CL11] S Cohen-Boulakia, U Leser, Search, adapt, and reuse: the future of scientific workflows, SIGMOD Rec., 40 (2), pp 6--16,2011

[CLM+15] V. Cuevas-Vicenttin, B. Ludäscheret, P. Missier, et al.: ProvONE: A PROV Extension Data Model for Scientific Workflow Provenance (2015) https://purl.dataone.org/provone-v1-dev

[CMH+20] R Celebi, R Moreira, A Hassan et al 2020. Towards FAIR protocols and workflows: the OpenPREDICT use case. PeerJ Computer Science 6:e281

[CMS+20] A Chapman, P. Missier, G. Simonelli, R. Torlone: Capturing and querying fine-grained provenance of preprocessing pipelines in data science. VLDB.144: 507-520 (2020)

[DCS09] D. De Roure, C Goble, R Stevens, The design and realisation of the myExperiment Virtual Research Environment for social sharing of workflows, Future Generation Comp. Syst., 2009, 25,pp 561-567

[DF08] SB Davidson, J Freire. Provenance and scientific workflows: challenges and opportunities. InProceedings of the 2008 ACM SIGMOD 2008 Jun 9 (pp. 1345-1350).

[DNN+21] W Digan, A Névéol, A Neuraz et al: Can reproducibility be improved in clinical natural language processing? A study of 7 clinical NLP suites. J. Am. Medical Informatics Assoc. 28(3): 504-515 (2021)

[DPL+14] ID Dinov, P Petrosyan, Z Liu et al High-throughput neuroimaging-genetics computational infrastructure. Frontiers in neuroinformatics. 2014 Apr 23;8:41.

[EPF+20] Ewels, Philip A., et al. "The nf-core framework for community-curated bioinformatics pipelines." Nature biotechnology 38.3 (2020): 276-278.

[EST+22] N. Etminan, DA de Sousa, C Tiseo et al.: European Stroke Organisation (ESO) guidelines on management of unruptured intracranial aneurysms. European Stroke Journal. 2022;7(3):LXXXI-CVI.

[GCS+20] Carole A. Goble, Sarah Cohen-Boulakia, Stian Soiland-Reyes et al.: FAIR Computational Workflows. Data Intell. 2(1-2): 108-121 (2020)

[GHZ+19] T Gao, X Han, H Zhuet al. 2019. FewRel 2.0: Towards More Challenging Few-Shot Relation Classification. In 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 6250–6255, Hong Kong, China.

[GG12] D. Garijo, Y. Gil: Augmenting PROV with Plans in P-PLAN: Scientific Processes as Linked Data. LISC@ISWC 2012

[GK+20] Groenen, Karlijn H. J. et al. “The de novo FAIRification process of a registry for vascular anomalies.” *Orphanet Journal of Rare Diseases* 16 (2020)

[GMC+21] Köhler, M. Gargano, N. Matentzoglu, L. C. Carmody, D. Lewis-Smith, N. A. Vasilevsky, et al., The Human Phenotype Ontology in 2021. Nucleic Acids Res. 49 (2021): D1207–D1217.

[GNT10] J Goecks, A Nekrutenko, J Taylor et al, Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences,
Genome Biol, vol 11(8), R86, 2010

[GRR+22] O Giraldo, M Ruano, R Richardsonet al (2022). Nanotate: Semantically Annotating Experimental Protocols with Nanopublications. Proc International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences (Vol. 3127, pp. 64-73). CEUR Workshop Proc. http://ceur-ws.org/Vol-3127/

[GSB20] A Gaignard, H Skaff-Molli, K Belhajjame. Findable and reusable workflow data products: A genomic workflow case study. Semantic Web, vol. 11, no. 5, pp. 751-763, 2020.

[HFF+22] N Hiebel, O Ferret, K Fort et al.: CLISTER : A Corpus for Semantic Textual Similarity in French Clinical Narratives. LREC 2022: 4306-4315
[HZD+22] X Han, W Zhao, N Ding et al. 2022. PTR: Prompt Tuning with Rules for Text Classification. AI Open, 3:182–192.
[IKJ+13] J. C. Ison, M. Kalas, I. Jonassen, et al. EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. Bioinform. 29(10): 1325-1332 (2013)
[LAB+06] B Ludäscher, I Altintas, C Berkley et al Scientific workflow management and the Kepler system. Concurr. Comput. Pract. Exp. 18(10): 1039-1065 (2006)
[LCG+22] M Louarn, F Chatonnet, X Garnier et al. Improving reusability along the data life cycle: a regulatory circuits case study. J Biomed Semant 13, 11 (2022).
[LNA+22] MA Levinson, J Niestroy, S Al Manir, et al (2022). FAIRSCAPE: a Framework for FAIR and Reproducible Biomedical Analytics. Neuroinformatics, 20(1), 187–202.
[LNF19] S Liu, KY Ngiam and M Feng 2019. Deep reinforcement learning for clinical decision support: a brief survey. arXiv preprint arXiv:1907.09475.
[LWY+22] Q Li, Y Wang, T You, et al. 2022. BioKnowPrompt: Incorporating imprecise knowledge into prompt-tuning verbalizer with biomedical text for relation extraction. Information Sciences, 617:346–358.
[LYF+22] P. Liu, W Yuan, J Fuet al. 2022. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. 2022. ACM Computer Survey.
[LYH+20] C Liang, Y Yu, H Jianget al 2020. BOND: BERT-Assisted Open-Domain Named Entity Recognition with Distant Supervision. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 1054–1064, New York, NY, USA. Association for Computing Machinery.
[LWL+22] T Li, Z Wang, W Lu et al 2022. Electronic health records based reinforcement learning for treatment optimizing. Information Systems, 104, p.101878.
[MBC13] P. Missier, K. Belhajjame, and J. Cheney. The W3C PROV family of specifications for modelling provenance metadata. Proc of the 16th International Conference on Extending Database Technology. 2013.
[MBC+14] L Murta, V Braganholo, F Chirigati et al: noWorkflow: Capturing and Analyzing Provenance of Scripts. IPAW 2014: 71-83
[MBR+09] M Mintz, S Bills, R Snowet al 2009. Distant supervision for relation extraction without labeled data. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pages 1003–1011.
[MBZ+08] P Missier, K Belhajjame, J Zhao, M Roos et al: Data Lineage Model for Taverna Workflows with Lightweight Annotation Requirements. IPAW2008: 17-30
[MGL+21] R Ma, T Gui, L Li, Q Zhanget al 2021. SENT: Sentence-level Distant Relation Extraction via Negative Training. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int Joint Conference on Natural Language Processing, pages 6201–6213, Online.
[MHM21] K Marzagão, TD Huynh, L Moreau (2021). Incremental Inference of Provenance Types. In: Provenance and Annotation of Data and Processes. IPAW 2021. LNCS, vol 12839. Springer
[NWL+11] Névéol A, Wilbur WJ, Lu Z. Extraction of data deposition statements from the literature: a method for automatically tracking research results. Bioinformatics. 2011. Dec 1;27(23):3306-12.
[Pel13] Peleg, M., 2013. Computer-interpretable clinical guidelines: a methodological review. Journal of biomedical informatics, 46(4), pp.744-763.
[POB+17] A Polyvyanyy, C Ouyang, A Barros et al Process querying: Enabling business intelligence through query-based process analytics. Decis. Support Syst. 100: 41-56 (2017)
[RCR+21] A Rogier, A Coulet, B Rance: Using an Ontological Representation of Chemotherapy Toxicities for Guiding Information Extraction and Integration from EHRs. MedInfo 2021: 91-95
[SBC+14] J. Starlinger, B. Brancotte, S. Cohen-Boulakia et al Similarity Search for Scientific Workflows, 2014, VLDB Endowment,7(12)
[SC18] A Smirnova, P. Cudré-Mauroux. 2018. Relation Extraction Using Distant Supervision: A Survey. ACM Computing Surveys, 51:106:1–106:35.

[SCL+12] J Starlinger, S Cohen-Boulakia, U Leser, (Re)Use in Public Scientific Workflow Repositories, Scientific and Statistical Database Management - SSDBM 2012, LNCS 7338, 361-378, Springer, 2012

[SCK+16]J Starlinger, S Cohen-Boulakia, S Khanna et al: Effective and efficient similarity search in scientific workflow repositories. Future Gener. Comput. Syst. 56: 584-594 (2016)

[SR+22] S Soiland-Reyes et al. (2022): Packaging research artefacts with RO-Crate. *Data Science* 5(2)

[SVR+19] SS Sahoo, J Valdez M Kim et al. ProvCaRe: characterizing scientific reproducibility of biomedical research studies using semantic provenance metadata. Int J of medical informatics. 2019 Jan 1;121:10-8.

[WDA+16] Wilkinson, Mark D., et al. "The FAIR Guiding Principles for scientific data management and stewardship." Scientific data 3.1 (2016): 1-9.

[WB13] Williams et al: Management of unruptured intracranial aneurysms, Neu Clin Pract 2013, 3 (2) 99-108

[WHF+13] K. Wolstencroft, R. Haines, D. Fellows et al. The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud, Nucleic acids research, gkt328, 2013

[WSL16] D Wiegandt, J Starlinger, U Leser: Graph n-grams for Scientific Workflow Similarity Search. LWDA 2016: 213-224

[WZA+20] Wei, Q., Y. Zhang, M. Amith, R. Lin, J. Lapeyrolerie, C. Tao, et H. Xu (2020). Recognizing software names in biomedical literature using machine learning. Health informatics journal 26(1), 21–33.