Découvrabilité et réutilisation de données produites par des workflows : un cas d'usage en génomique

Alban Gaignard, Hala Skaf-Molli, Khalid Belhajjame

Découvrabilité et réutilisation de données produites par des workflows : un cas d'usage en génomique

Alban Gaignard[1], Hala Skaf-Molli[2], Khalid Belhajjame[3]

[1] l'institut du thorax, INSERM, CNRS, University of Nantes, Nantes, France
[2] LS2N, University of Nantes, Nantes, France
[3] LAMSADE, PSL, Université Paris-Dauphine, Paris, France

alban.gaignard@univ-nantes.fr
hala.skaf@univ-nantes.fr
kbelhajj@googlemail.com

Résumé

Les systèmes de workflows ont largement contribué à améliorer la reproductibilité des expériences scientifiques. Cependant, relativement peu de travaux ont porté sur la réutilisation des données produites au cours de l'exécution. Dans cet article, nous faisons l'hypothèse que ces données intermédiaires doivent être considérées comme des objets de premier ordre, qui doivent être conservés et publiés. Non seulement cela permettra d'économiser des ressources de calcul et de stockage, mais surtout cela facilitera et accélérera l'évaluation de nouvelles hypothèses. Pour aider les scientifiques à annoter ces données, nous exploitons plusieurs sources d'information : i) les informations de provenance capturées lors de l'exécution des workflows, et ii) les annotations de domaine qui sont fournies par des catalogues sémantiques d'outils, tels que Bio.Tools. Finalement, nous montrons, sur un scénario réel de bioinformatique, comment des graphes de provenance peuvent être transformés et résumés, à destination des utilisateurs et des machines.

Mots-clés

FAIR, reproductibilité, workflows scientifiques.

Abstract

Workflow systems have played an important role in facilitating the reproducibility of scientific experiments, yet, little work has been devoted to enhance the reuse of produced data. We argue that these intermediate data should be considered as first-order objects, which are worthy of preservation and publication. Not only will this save computational resources, but more importantly it will ease and accelerate the evaluation of new hypotheses. To help scientists annotate such produced data, we exploit multiple sources of information : i) provenance information captured during the execution of workflows, and ii) domain annotations provided by semantic catalogs of tools, such as Bio.Tools. Finally, we show, on a real bioinformatics scenario, how provenance graphs can be transformed and synthesized, for human and machine use.

Keywords

FAIR, reproducibility, scientific workflows.

1 Introduction

Les sciences dirigées par les données ont amené ces dernières années un changement de paradigme. L'évaluation d'hypothèses scientifiques repose de plus en plus sur des codes informatiques d'analyse, organisés sous la forme de *pipelines* (ou *workflows*), et exécutés sur des masses de données[1, 21, 2, 11]).

Pour répondre aux enjeux de reproductibilité [25], les scientifiques ont été encouragés à ne pas seulement rendre compte de leurs résultats, mais aussi à documenter leurs méthodes et expériences, ainsi que l'ensemble des données analysées et produites. Un certain nombre de méthodes et d'outils ont été proposés pour aider les scientifiques dans cette tâche [9, 16, 3].

Malgré l'intérêt de ces propositions visant à faciliter la répétabilité des expériences, elles n'apportent pas encore de réponse quant à la réutilisation des données produites.

Nous faisons l'hypothèse que toutes les données produites par les *workflows* associés aux expériences doivent être considérées comme des objets de premier ordre, afin d'être plus facilement découvrables, accessibles et finalement réutilisables par les membres de la communauté scientifique.

Dans cet article, nous montrons comment nous pouvons combiner les métadonnées de provenance avec des connaissances externes associées aux *workflows* et aux outils bioinformatiques (Bio.Tools [19]) pour promouvoir le partage et la réutilisation des données traitées. **Notre objectif principal est de promouvoir la réutilisation des données traitées afin de limiter la duplication des efforts de calcul et de stockage associés à la ré-exécution de *workflows***.

Les contributions de cet article sont les suivantes :

— un scénario concret de réutilisation de données produites par des *workflows* dans le domaine de la bioinformatique,
— une approche basée sur des graphes de connais-

sances pour l'annotation sémantique des données brutes,
— une évaluation expérimentale de l'approche à l'aide d'un *workflow* réel en bioinformatique.

Cet article est organisé comme suit. La section 2 présente les motivations et définit le problème scientifique. La section 3 détaille l'approche FRESH proposée. La section 4 présente nos résultats expérimentaux. La section 5 résume les travaux de l'état de l'art. Enfin, les conclusions et les travaux futurs sont exposés dans la section 6. Les lecteurs intéressés peuvent accéder à la version étendue de ce travail qui a été publiée dans un journal de langue anglaise [15].

2 Motivations et problématique

Nous motivons notre proposition à partir d'un *workflow* de séquençage d'exomes. Il consiste (1) à aligner les séquences d'ADN codantes d'un échantillon sur un génome de référence et (2) à détecter leurs mutations génétiques. La figure 1 résume les tâches d'analyse bioinformatique. Pour des raisons de clarté, nous masquons dans ce scénario certaines des étapes de traitement mineures telles que le tri des bases d'ADN.

Dans des conditions réelles, de telles analyses nécessitent beaucoup de temps de calcul et de capacité de stockage. A titre d'exemple, des *workflows* similaires sont exécutés en production au centre national de séquençage (CNRGH). Pour un échantillon typique en séquençage d'exome (9,7 Go compressé), 18,6 Go sont nécessaires pour stocker les données compressées d'entrée et de sortie. Sur une infrastructure de calcul *HPC* (7 nœuds de calculs avec 28 cœurs Intel Broadwell), 2 heures et 27 minutes sont nécessaires pour produire un fichier de variants VCF annoté, ce qui correspondrait sur un seul CPU à 158 heures cumulées pour un seul échantillon, soit 6 jours de calcul sur un seul CPU.

Nous faisons donc l'hypothèse que la réutilisation des données déjà analysées, est essentielle pour accélérer la recherche sur des sujets similaires ou connexes. Dans le *workflow* de la figure 1, `GRCh37` est considéré comme hautement réutilisable car il constitue un atlas de référence pour les séquences génomiques humaines, et il résulte de l'état des connaissances scientifiques à un moment donné. Par ailleurs, les fichiers `BAM` peuvent également être considérés comme plus réutilisables que les données brutes car ils ont été alignés sur cet atlas et bénéficient donc des connaissances associées à cet atlas génomique. Par exemple, ils fournissent la relation entre les séquences et les gènes connus, ils peuvent être visualisés à l'échelle du génome, ou bien réutilisé pour générer des séquences brutes non alignées.

Pour les scientifiques il est très difficile de répondre à des questions telles que "puis-je réutiliser ces fichiers dans le contexte de mes travaux ?". Dans cet exemple, si l'on souhaite réutiliser le fichier de variants final, il faut absolument connaître la version du génome de référence ainsi que le contexte scientifique de l'étude, les phénotypes associés aux échantillons, ainsi que les relations possibles entre échantillons. Enfin, il est également essentiel de disposer d'informations précises sur l'algorithme de détection de variants en raison des seuils de détection internes [22]. Plus généralement, il faut non seulement des informations détaillées de provenance concernant l'historique des traitements de données, mais également des annotations de domaine basées sur des vocabulaires contrôlés (problème 1). Ces vocabulaires existent mais l'annotation des données traitées avec des concepts spécifiques à un domaine demande beaucoup de temps et d'expertise (problème 2).

Dans ce travail, nous montrons comment améliorer la réutilisation des données (intermédiaires) de *workflows* en tirant parti (1) des efforts de la communauté visant à cataloguer sémantiquement les outils en bioinformatique, et (2) des capacités d'automatisation et de capture d'information de provenance des systèmes de gestion de *workflows* pour automatiser l'annotation des données traitées.

3 Approche

FRESH est une approche visant à améliorer la découvrabilité (*Findability*) et la réutilisation (*Reusability*) des données produite et analysées par des *workflows* dans le domaine de la génomique. Les principes FAIR [26, 27] et les approches Linked data [6, 5] constituent les piliers conceptuels et technologiques de cette démarche.

Nous abordons la question de la découvrabilité en nous appuyant sur les données liées sur le web (*Linked Data*), à savoir l'association d'un URI à chaque entité, la mise en relation de ces entités sous la forme de graphes de connaissances RDF, et l'utilisation de vocabulaires contrôlés qui définissent la nature de ces entités et leurs relations.

Très liée au contexte scientifique, la réutilisation des données est plus difficile. Des éléments de réponse ont été proposés pour le partage FAIR des données génomiques [12], cependant, proposer et évaluer la réutilisabilité des données est toujours un défi et un travail en cours [28]. Dans ce travail, nous nous concentrons sur les données réutilisables comme annotées avec des informations suffisamment complètes permettant une meilleure traçabilité, interprétabilité à la fois par des utilisateurs ou des machines.

Pour une meilleure traçabilité, les graphes de provenance sont nécessaires afin de suivre le processus de génération des données.

Pour une meilleure interprétabilité, des informations contextuelles [26] sont nécessaires, par exemple : (i) les hypothèses de recherche, les laboratoires de recherche, les conditions expérimentales, les résultats antérieurs (publications scientifiques), et (ii) le contexte technique en termes de matériel, de méthodes, de sources de données, de logiciels utilisés (algorithmes, requêtes). Ces données doivent être annotées avec des vocabulaires spécifiques d'un domaine. Pour expliciter les connaissances associées aux étapes d'analyse de données, nous pouvons nous appuyer sur l'ontologie EDAM [1] qui est activement développée et utilisée dans le cadre du registre d'outils bioinformatiques

1. `http://edamontology.org`

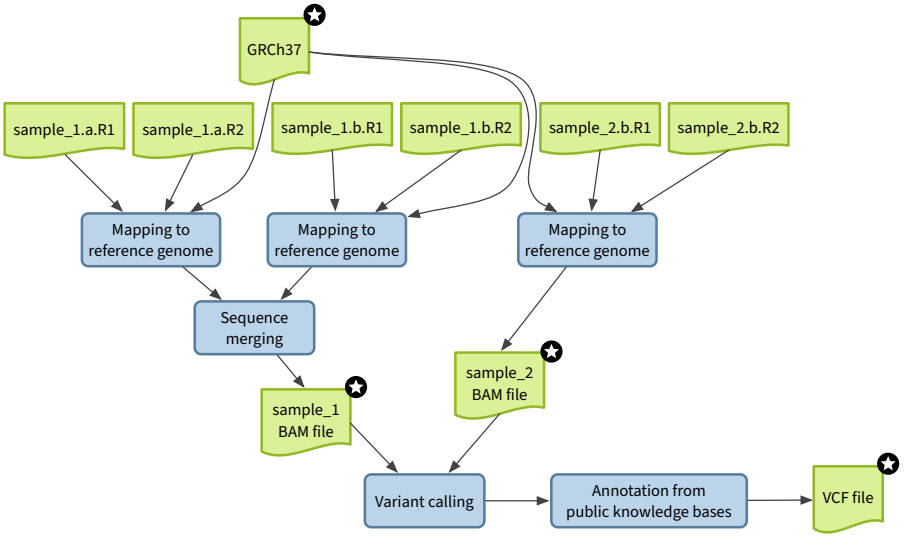

FIGURE 1 – Un *workflow* de bioinformatique typique visant à détecter et annoter des mutations génétiques. Les données sont en vert et les étapes de traitement sont en bleu.

Bio.Tools [2], et qui organise les concepts et relations sémantiques dans le domaine de la bioinformatique. Cependant, cette ontologie ne permet pas de décrire le contexte scientifique associé à un *workflow*. Pour résoudre ce problème, nous nous appuyons sur l'ontologie *Micropublications* [10] qui a été proposée pour représenter formellement les approches scientifiques, les hypothèses, ou les éléments de preuve, dans la perspective de faciliter l'exploitation des articles scientifiques par des algorithmes.

La figure 2 illustre notre approche pour améliorer la réutilisabilité des données. La première étape consiste à capturer la provenance pour toutes les exécutions d'un *workflow*. PROV-O [3] est le standard *de facto* pour décrire et échanger des graphes de provenance. Bien que la capture de la provenance puisse être facilement gérée dans les moteurs de *workflows*, il n'existe pas de moyen systématique de relier une activité PROV-O (l'exécution réelle d'un outil) à l'agent logiciel correspondant *i.e.* Le logiciel responsable du traitement des données). Pour résoudre ce problème, nous proposons de fournir, au moment de la conception du workflow, l'identifiant de l'outil dans le catalogue des outils. Cela permet de générer une trace de provenance qui associe chaque exécution (`prov:wasAssociatedWith`), et donc chaque donnée consommée et produite, à l'identifiant du logiciel.

Ensuite, nous assemblons un graphe de connaissances bioinformatiques qui relie (1) les annotations des outils, recueillies dans le registre Bio.Tools, fournissant des informations sur les fonctions des outils (opérations bioinformatiques EDAM) et le type de données qu'ils consomment et produisent, (2) l'ontologie EDAM complète, pour accéder par exemple aux définitions et synonymes, (3) le graphe PROV-O résultant de l'exécution d'un *workflow* qui fournit des métadonnées de provenance techniques et génériques, et (4) le contexte expérimental en utilisant les micropublications pour décrire les questions et hypothèses scientifiques associées à l'expérience.

Enfin, sur la base de requêtes de provenance spécifiques au domaine, la dernière étape consiste à extraire quelques données significatives du graphe de connaissances, afin de fournir aux scientifiques des résultats intermédiaires ou finaux plus réutilisables, et de fournir des historiques de données découvrables et interrogeables par les machines.

Dans le reste de cette section, nous nous appuyons sur le langage de requête SPARQL pour interagir avec le graphe de connaissances en termes d'extraction et d'enrichissement des connaissances.

```
SELECT ?d_label ?title ?f_def ?st WHERE {
    ?d rdf:type prov:Entity ;
       prov:wasGeneratedBy ?exec ;
       rdfs:label ?d_label .

    ?exec prov:wasAssociatedWith ?tool ;
          prov:wasStartedBy ?wf .

    ?tool dc:title ?title ;
       biotools:has_function ?f .

    ?f rdfs:label ?f_label ;
       oboInOwl:hasDefinition ?f_def .

    ?wf mp:supports ?c .
    ?c rdf:type mp:Claim ;
       mp:statement ?st .
}
```

Requête 1 – Requête SPARQL permettant de lier les données produite aux outils/algorithmes.

La requête 1 vise à lier des données avec la définition de l'opération bioinformatique dont elles résultent. Dans cette requête SPARQL, nous identifions d'abord les données (`prov:Entity`), l'exécution de l'outil dont elles résultent (`prov:wasGeneratedBy`), et le logiciel utilisé (`prov:wasAssociatedWith`). Ensuite, nous obtenons du

2. http://bio.tools
3. https://www.w3.org/TR/prov-o/

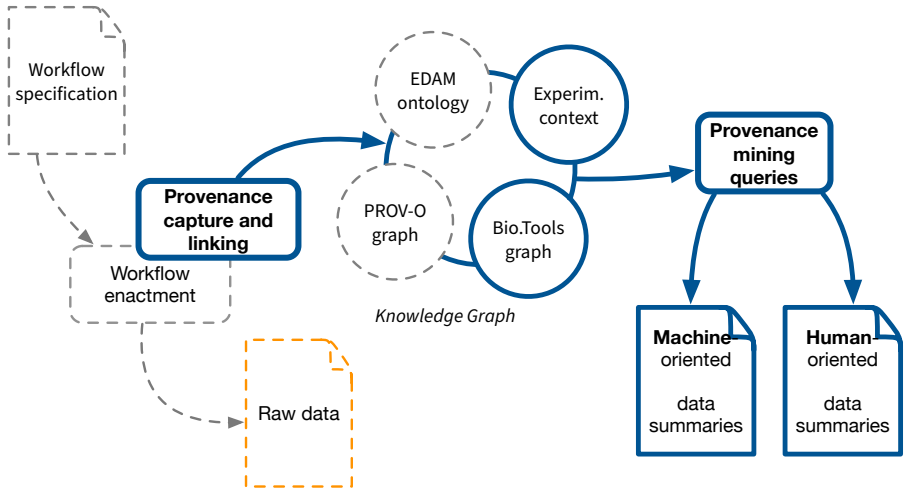

FIGURE 2 – Base de connaissances produite à partir des informations de provenance et des annotations des outils afin d'automatiser la production de résumés destinés aux utilisateurs et aux machines.

sous-graphe Bio.Tools l'annotation EDAM qui spécifie la fonction de l'outil (`biotools:has_function`). La définition de la fonction de l'outil est obtenue dans l'ontologie EDAM (`oboInOwl:hasDefinition`). Enfin, nous identifions le contexte scientifique de l'expérience en faisant correspondre les déclarations exprimées en langage naturel (`mp:Claim`, `mp:statement`).

La requête 2 montre comment un motif de provenance peut-être transformé pour fournir un résumé des principales étapes de traitement, en s'appuyant sur une ontologie de domaine. L'idée consiste d'abord à identifier tous les liens de dérivation de données (`prov:wasDerivedFrom`). Ensuite les exécutions d'outils sont identifiées ainsi que les agents logiciels correspondants, et la fonctionnalité des outils. Nous exploitons la propriété `biotools:has_function`. Une fois ce patron identifié, un nouveau graphe est créé à l'aide d'une clause CONSTRUCT. Il représente une chaîne ordonnée d'étapes de traitement (`p-plan:wasPreceededBy`).

```
CONSTRUCT {
    ?x2 p-plan:wasPreceededBy ?x1 .
    ?x2 prov:wasAssociatedWith ?t2 .
    ?x1 prov:wasAssociatedWith ?t1 .
    ?t1 biotools:has_function ?f1 .
    ?f1 rdfs:label ?f1_label .
    ?t2 biotools:has_function ?f2 .
    ?f2 rdfs:label ?f2_label .
} WHERE {
    ?d2 prov:wasDerivedFrom ?d1 .

    ?d2 prov:wasGeneratedBy ?x2 ;
        prov:wasAssociatedWith ?t2 ;
        rdfs:label ?d2_label .

    ?d1 prov:wasGeneratedBy ?x1 ;
        prov:wasAssociatedWith ?t1 ;
        rdfs:label ?d1_label .

    ?t1 biotools:has_function ?f1 .
    ?f1 rdfs:label ?f1_label .

    ?t2 biotools:has_function ?f2 .
    ?f2 rdfs:label ?f2_label .
}
```

Requête 2 – Requête SPARQL permettant de produire un *workflow* abstrait.

4 Résultats expérimentaux

4.1 Graphes de provenance

Nous avons expérimenté notre approche sur un *workflow* de séquençage d'exome [4], conçu et exploité par la plateforme de génomique et de bioinformatique GenoBird. Il met en œuvre le scénario de motivation que nous avons présenté dans la section 2. Nous supposons que, sur la base de l'approche présentée précédemment, le *workflow* a été exécuté, la provenance associée a été capturée et le graphe de connaissances a été assemblé.

Le graphe de provenance consiste en un graphe RDF avec 555 triplets exploitant l'ontologie PROV-O.

L'interprétation de ce graphe de provenance est difficile d'un point de vue humain en raison du nombre de noeuds et d'arêtes et, surtout, de l'absence de termes spécifiques à un domaine.

4.2 Résumés de données destinés aux utilisateurs

Sur la base de la requête 1 et d'un modèle textuel, nous montrons dans la figure 3 des phrases qui ont été générées automatiquement à partir du graphe de connaissances. Elles ont pour but de fournir aux scientifiques des informations explicites sur la façon dont les données ont été produites, et sur leur contexte scientifique, en utilisant des termes spécifiques au domaine.

Les procédures complexes d'analyse des données nécessiteraient un long texte et de nombreuses articulations logiques pour être compréhensibles. Les diagrammes visuels fournissent une représentation compacte pour le traitement de données complexes et constituent donc un moyen intéressant d'assembler des résumés de données pour les scientifiques.

4. https://gitlab.univ-nantes.fr/bird_pipeline_registry/exome-pipeline

```
The file <VCF/hapcaller.recal.combined.annot.
gnomad.vcf.gz> results from tool
<gatk2_variant_annotator-IP> which <Predict the
effect or function of an individual single
nucleotide polymorphism (SNP).>
It was produced in the context of <Rare Coding
Variants in ANGPTL6 Are Associated with Familial
Forms of Intracranial Aneurysm>
```

FIGURE 3 – Résumé textuel basé sur l'ontologie EDAM.

TABLE 1 – Temps de calcul pour la génération des résumés de données

Graphe RDF	chargement	résumés text.	NanoPub.	Visu.
218 906 triplets	22.7s	1.2s	61ms	1.5s

Un exemple de diagramme récapitulatif est fourni par la Figure 4. Les flèches noires représentent le flux logique du traitement de données, et les ellipses noires représentent la nature du traitement de données, en termes d'opérations EDAM. Le diagramme montre que les fichiers en bleu résultent d'une étape de traitement effectuant une opération de type *"SNP annotation"*, telle que définie dans l'ontologie EDAM.

Ces visualisations fournissent aux scientifiques les moyens de positionner un résultat intermédiaire, par exemple des séquences génomiques alignées sur un génome de référence (fichier BAM), ou des variants génomiques (fichier VCF) dans le contexte d'un processus complexe d'analyse. Alors qu'un bioinformaticien expert n'aura pas besoin de ces résumés, nous considérons que visualiser et rendre explicites ces résumés est d'un intérêt majeur pour mieux réutiliser les données scientifiques, voire fournir un premier niveau d'explication en termes de concepts spécifiques au domaine d'application.

4.3 Résumés de données destinés aux machines

Les principes *Linked Data* préconisent l'utilisation de vocabulaires contrôlés et d'ontologies pour fournir des connaissances lisibles par l'homme et par la machine. Nous montrons dans la figure 5 comment des annotations spécifiques au domaine (EDAM), peuvent être agrégées et partagées entre des machines en exploitant le vocabulaire NanoPublication. Ces résumés peuvent être indexés et découverts sémantiquement, conformément aux principes de *Findability* de FAIR.

4.4 Implémentation

Nous avons légèrement étendu le moteur de workflow Snakemake [20] avec un module de capture de provenance [5]. Nous avons également développé un *crawler* [6] qui construit un jeu de données RDF à partir du registre Bio.Tools. Les résultats présentés dans la section 4 ont été obtenus en exécutant un Notebook Jupyter [7].

5. https://github.com/albangaignard/snakemake/tree/research-objects

6. https://github.com/bio-tools/biotoolsRdf

7. https://github.com/albangaignard/fresh-toolbox

Nous avons simulé l'exécution du *workflow* d'analyse de données d'exome pour évaluer le temps de calcul des résumés de données, à partir d'un graphe de connaissances RDF. Cette simulation a permis de ne pas être impacté par les temps de calcul réels de l'analyse des données génomiques. Le tableau 1 décrit les temps de calcul en utilisant un ordinateur portabe MacBook Pro Core i5 de 16 Go et 2,9 GHz. Nous avons mesuré 22,7 secondes pour charger en mémoire le graphe de connaissances complet (218 906 triplets) décrivant l'exécution du *workflow* via son graphe de provenance, le registre d'outils Bio.Tools et l'ontologie EDAM. Les résumés de données textuels ont été obtenus en 1.2s, la *NanoPublication* a été générée en 61 ms, et enfin il a fallu 1.5 s pour générer une visualisation sous la forme de graphe du résumé. Ce sur-coût peut être considérée comme négligeable par rapport aux ressources informatiques nécessaires pour analyser des données de séquençage d'exome, comme indiqué dans la section 2.

4.5 Discussion

Les résultats expérimentaux montrent qu'il est possible de générer des résumés de données qui fournissent des informations précieuses sur les données du *workflow*. Nous nous concentrons sur les annotations spécifiques au domaine afin de promouvoir la découvrabilité et la réutilisation des données, en lien avec les principes FAIR [8] avec une attention particulière pour les *workflows* en génomique.

Pour ce qui est de la découvrabilité, FRESH répond en partie aux exigences F1 ((Les (méta)données ont un identifiant unique et persistant), F2 (Les données sont décrites avec des métadonnées riches) et F3 (Les métadonnées incluent clairement et explicitement l'identifiant des données qu'elles décrivent) car (i) nous attribuons des identifiants uniques universels (UUID) aux entités de provenance et (ii) nous réutilisons les ontologies NanoPublication et EDAM pour partager et réutiliser les données produites. Les nanopublications générées pourraient être publiées soit via un serveur SPARQL, soit par le réseau de serveurs NanoPublication.

Pour la réutilisabilité, FRESH addresse R1.2 (les (méta)données sont associées à des informations de provenance détaillées) et R1.3 (les (méta)données répondent aux standards adoptés par les communautés). Comme illustré dans les sections précédentes, FRESH peut être utilisé pour générer des résumés de données destinés aux scientifiques ou aux machines.

Toujours dans le contexte de l'analyse des données génomiques, un scénario de réutilisation typique consisterait à exploiter les variantes génomiques annotées, pour effectuer une analyse statistique des variantes rares. Si l'on considère qu'aucune sémantique n'est attachée aux noms de fichiers ou d'outils, les informations de provenance générique ne permettraient pas de fournir des informations sur la nature du traitement de données. En regardant le diagramme destinés aux utilisateurs, ou en laissant un algorithme interroger la nanopublication, destinée aux machines, produite par FRESH, les scientifiques seraient plus facilement en mesure de comprendre que le fichier résulte d'une annotation

8. https://www.go-fair.org/fair-principles

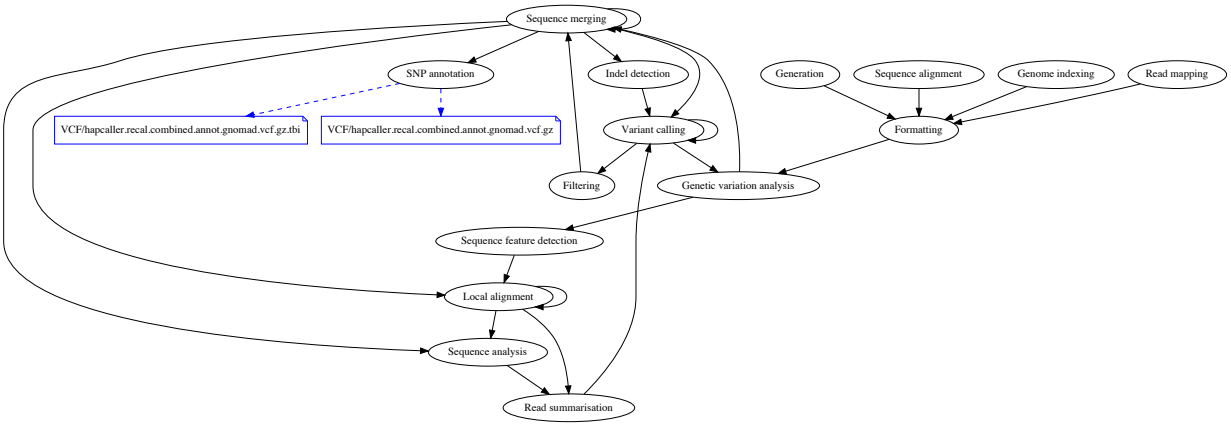

FIGURE 4 – Un diagramme compilé automatiquement à partir du graphe de provenance et de connaissances spécifiques du domaine et destiné aux scientifiques.

```
:head {
    _:np1 a np:Nanopublication .
    _:np1 np:hasAssertion :assertion .
    _:np1 np:hasProvenance :provenance .
    _:np1 np:hasPublicationInfo :pubInfo .
}

:assertion {
  <http://snakemake-provenance/Samples/Sample1/
  BAM/Sample1.merged.bai> rdfs:seeAlso
  <http://edamontology.org/operation_3197> .

  <http://snakemake-provenance/VCF/hapcaller.
  indel.recal.filter.vcf.gz> rdfs:seeAlso
  <http://edamontology.org/operation_3695> .
}
```

FIGURE 5 – Extrait d'une *NanoPublication* permettant d'aggréger des assertions spécifiques du domaine, des informations de provenance et de publication.

de polymorphismes génétiques (SNP) qui a été précédée d'une étape de détection de variants elle-même précédée d'une étape de détection d'insertion/délétion (Indel).

Nous nous sommes concentrés dans ce travail sur le domaine de la bioinformatique et avons exploité Bio.Tools, un effort communautaire à grande échelle visant à cataloguer sémantiquement les algorithmes/outils disponibles. Dès que des catalogues d'outils sémantiques seront disponibles pour d'autres domaines, FRESH pourra être appliqué afin d'améliorer la découvrabilité et la réutilisation des données traitées. Même s'ils sont plus récents, des efforts similaires s'adressent à la communauté de la bioimagerie grâce à la mise en place du registre de bioimagerie BISE (projet européen COST Neubias), et l'extension de l'ontologie EDAM pour la bioimagerie.

Dans ce travail, nous avons validé notre solution manuellement dans le cadre d'un *workflow* réel de génomique. Un ensemble de données et de workflows de référence permettrait de plus facilement évaluer les approches visant à améliorer la réutilisation de données scientifiques. Ces données et workflows de référence pourraient résulter des activités de différentes communautés scientifiques afin de mieux répondre aux enjeux des sciences ouvertes et reproductibles.

5 Etat de l'art

Notre approche est liée aux travaux visant à faciliter la préservation, la reproductibilité et la réutilisation des ressources scientifiques. OBI (Ontology for Biomedical Investigations) [7] et le modèle ISA (Investigation, Study, Assay) [23] sont deux modèles largement utilisés dans le domaine des sciences de la vie pour décrire des travaux scientifiques. Research Objects [4] propose une suite d'ontologies visant à agréger des spécifications de *workflows*, leurs exécutions et leur contexte scientifique. ReproZip [9] est une autre solution qui permet de créer des archives comprenant les dépendances nécessaires pour reproduire un *workflows*.

Les solutions ci-dessus aident les scientifiques à agréger leurs informations dans un seul conteneur. Cependant, elles n'aident pas pour l'annotation des résultats d'expériences. Pour ce problème, Alper *et al.* [3] et Gaignard *et al.* [16] ont développé des solutions qui permettent de dériver des annotations à partir des *workflows* et de les résumer.

L'ontologie PROV-O, recommandation du W3C et ses extensions, ProvONE [9], OPMW [10], Wfprov [11] ou P-Plan [17], présentent un intérêt particulier pour notre travail. Elles permettent de poser des questions sur le "pourquoi" et le "comment" de la production des données. Cependant, répondre à des questions telles que "ces données sont-elles utiles pour mon expérience ?" ou bien "sont-elles de qualité suffisante ?" est difficile et nécessite des annotations spécifiques d'un domaine, non couvertes par les modèles de provenance génériques.

Dans nos travaux précédents [16], nous avons proposé *PoeM* une approche pour générer des rapports d'expérience basés sur la provenance et les annotations des utilisateurs.

9. https://purl.dataone.org/provone-v1-dev
10. https://www.opmw.org
11. http://purl.org/wf4ever/wfprov#

SHARP [13, 14] étend *PoeM* pour les *workflows* s'exécutant sur différents systèmes et produisant des traces de provenance hétérogènes. Dans ce travail, nous nous appuyons sur ces travaux pour annoter et résumer les informations de provenance, en nous concentrant sur les données plutôt que sur le *workflow* lui-même.

La proposition de Garijo et Gil [18] est peut-être la plus proche de la nôtre dans le sens où elle se concentre sur les données, et génère des textes à partir des informations de provenance. Dans ce travail, nous nous concentrons plutôt sur l'annotation des données intermédiaires du *workflow*.

[24] vise à identifier les similitudes entre les *workflows* à partir de leur structure, des noms de leurs modules, et des auteurs associés. Notre objectif est différent dans la mesure où nous voulons promouvoir la réutilisation non seulement des *workflows*, mais aussi des données produites, en nous appuyant sur des techniques de résumé.

Cerezo et al. [8] ont proposé un modèle de *workflow* conceptuel, proche du domaine d'expertise de l'utilisateur final, visant à améliorer le partage et la réutilisation des *workflows* scientifiques. Dans notre approche, nous nous concentrons sur la réutilisation des données intermédiaires produites alors que Cerezo *et al.* se concentrent sur la réutilisation du processus de transformation de données lui-même. De plus, notre approche est basée sur l'exécution des *workflows*, et tend à limiter la sollicitation d'experts du domaine, en exploitant des catalogues d'outils déjà annotés sémantiquement.

Notre travail est également lié aux efforts de la communauté scientifique pour créer des entrepôts ouverts pour la publication de données scientifiques. Par exemple, Figshare [12] et Dataverse [13], qui aident les institutions universitaires à stocker, partager et gérer tous les résultats de leurs recherches. Les résumés de données que nous produisons peuvent être publiés dans ces entrepôts. Cependant, nous pensons que les résumés que nous produisons sont mieux adaptés aux entrepôts qui publient des graphes de connaissances, par exemple celui créé par le projet whyis (http ://tetherless-world.github.io/whyis/).

6 Conclusion et perspectives

Dans cet article, nous proposons une approche visant à rendre les données des *workflows* scientifiques plus facilement découvrables et réutilisables, à partir d'un exemple dans le domaine de la génomique. Pour cela, nous générons des résumés de données, à partir de métadonnées de provenance et d'une ontologie en bioinformatique. FRESH permet de produire des résumés de données concis pour les scientifiques et pour les machine. Les résultats expérimentaux montrent l'efficacité de FRESH en termes de temps de calcul, négligeable par rapport aux ressources informatiques requises pour analyser les données de génomique. Afin d'évaluer notre approche, nous souhaiterions mener une étude auprès des plate-formes de bioinformatique fédérées dans le cadre de l'infrastructure nationale de recherche IFB. Ces plate-formes développent et exécutent à grande échelle des *workflows* d'analyse de données dans le domaine de la génomique. Cette communauté de bioinformaticiens fait face aux enjeux de reproductibilité des analyses, de partage et de réutilisation des données, et pourra permettre d'évaluer la pertinence des résumés de données introduits dans FRESH. Cependant, mettre en place une telle étude nécessite des développements logiciels *open source* pour intégrer la capture de méta-données de provenance dans des moteurs de workflows utilisés en routine tels que SnakeMake, NextFlow ou encore Galaxy.

Dans un contexte de sciences ouvertes et reproductibles, il nous parait critique d'inciter les scientifiques (1) à produire des annotations sémantiques de haute qualité pour décrire les workflows et algorithmes (2) à produire des résumés de données sémantiques et inter-opérables afin de promouvoir la découvrabilité et la réutilisation des données scientifiques.

12. https://figshare.com/
13. https://dataverse.org/

Remerciements

Nous remercions la plateforme BiRD (Biogenouest, IFB) pour son soutien technique et l'utilisation de son infrastructure.

Références

[1] G. R. Abecasis, A. Auton, B., et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422) :56–65, 2012.

[2] P. Alper. *Towards harnessing computational workflow provenance for experiment reporting*. PhD thesis, University of Manchester, UK, 2016.

[3] P. Alper, K. Belhajjame, et al. Automatic versus manual provenance abstractions : Mind the gap. In *8th USENIX Workshop on the Theory and Practice of Provenance, TaPP 2016, Washington, D.C., USA, June 8-9, 2016.* USENIX, 2016.

[4] K. Belhajjame, J. Zhao, et al. Using a suite of ontologies for preserving workflow-centric research objects. *J. Web Semant.*, 32 :16–42, 2015.

[5] C. Bizer, T. Heath, et al. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3) :1–22, 2009.

[6] C. Bizer, Vidal M.-E., and H. Skaf-Molli. Linked open data. In Ling Liu and M. Tamer Özsu, editors, *Encyclopedia of Database Systems*. Springer, 2017.

[7] R. R. Brinkman, M. Courtot, et al. Modeling biomedical experimental processes with obi. In *Journal of biomedical semantics*, volume 1, page S7. BioMed Central, 2010.

[8] N. Cerezo and J. Montagnat. Scientific workflow reuse through conceptual workflows on the virtual imaging platform. In *Proceedings of the 6th Workshop on Workflows in Support of Large-scale Science*, WORKS '11, pages 1–10, New York, NY, USA, 2011. ACM.

[9] F. Chirigati, R. Rampin, et al. Reprozip : Computational reproducibility with ease. In *Proceedings of the 2016 International Conference on Management of Data*, pages 2085–2088. ACM, 2016.

[10] T. Clark, P. N. Ciccarese, et al. Micropublications : A semantic model for claims, evidence, arguments and annotations in biomedical communications. *Journal of Biomedical Semantics*, 2014.

[11] S. Cohen Boulakia, K. Belhajjame, et al. Scientific workflows for computational reproducibility in the life sciences : Status, challenges and opportunities. *Future Generation Comp. Syst.*, 75 :284–298, 2017.

[12] M. Corpas, N. V. Kovalevskaya, et al. A fair guide for data providers to maximise sharing of human genomic data. *PLoS Computational Biology*, 14(3), 2018.

[13] A. Gaignard, K. Belhajjame, et al. SHARP : harmonizing and bridging cross-workflow provenance. In *The Semantic Web : ESWC 2017 Satellite Events - ESWC 2017 Satellite Events, Portorož, Slovenia,Revised Selected Papers*, pages 219–234, 2017.

[14] A. Gaignard, K. Belhajjame, et al. SHARP : harmonizing cross-workflow provenance. In *Proceedings of the Workshop on Semantic Web Solutions for Large-scale Biomedical Data Analytics co-located with 14th Extended Semantic Web Conference, SeWeBMeDA@ESWC 2017, Portoroz, Slovenia.*, pages 50–64, 2017.

[15] A. Gaignard, H. Skaf-Molli, and K. Belhajjame. Findable and reusable workflow data products : A genomic workflow case study. *Semantic Web*, 11(5) :751–763, 2020.

[16] A Gaignard, H. Skaf-Molli, et al. From scientific workflow patterns to 5-star linked open data. In *8th USENIX Workshop on the Theory and Practice of Provenance*, 2016.

[17] D. Garijo and Y Gil. Augmenting PROV with plans in PPLAN : scientific processes as linked data. In *Proceedings of the Second International Workshop on Linked Science 2012 - Tackling Big Data, Boston, MA, USA, November 12, 2012*. CEUR-WS.org, 2012.

[18] Y. Gil and D. Garijo. Towards automating data narratives. In *Proceedings of the 22Nd International Conference on Intelligent User Interfaces*, IUI '17, pages 565–576, New York, NY, USA, 2017. ACM.

[19] J. Ison, K. Rapacki, et al. Tools and data services registry : A community effort to document bioinformatics resources. *Nucleic Acids Research*, 2016.

[20] J. Köster and S. Rahmann. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19) :2520–2522, 2012.

[21] J. Liu, E. Pacitti, et al. A survey of data-intensive scientific workflow management. *Journal of Grid Computing*, 13(4) :457–493, 2015.

[22] N. D. Olson, S. P. Lund, et al. Best practices for evaluating single nucleotide variant calling methods for microbial genomics, 2015.

[23] P. Rocca-Serra, M Brandizi, et al. Isa software suite : supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics*, 26(18) :2354–2356, 2010.

[24] J. Starlinger, S. Cohen-Boulakia, and U. Leser. (Re)use in public scientific workflow repositories. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012.

[25] V. Stodden. The scientific method in practice : Reproducibility in the computational sciences. *SSRN Electronic Journal*, 2010.

[26] M. D. Wilkinson et al. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3, 2016.

[27] M. D. Wilkinson et al. A design framework and exemplar metrics for fairness. *Scientific Data*, 5, 2018.

[28] M. D. Wilkinson et al. Fairmetrics/metrics : Proposed fair metrics and results of the metrics evaluation questionnaire. 2018.