Inserm

La science pour la santé
From science to health

2ème réunion annuelle du noeud RDA France
Paris, 12 septembre 2019
Atelier Données de santé / Provenance & Reproductibilité

Programme de l'atelier

10h00 : Présentation / Introduction (Isabelle Perseil)

10h15 : Mireille Louys

Métadonnées pour la provenance :modèle et exemples d'implémentation pour les données d'observations en Astrophysique

10h45 : Sarah Cohen-Boulakia

Reproductibilité computationnelle en sciences de la vie et workflows scientifiques : état-des lieux et Opportunités

11h30 : Alban Gaignard et Khalid Belhajjame

Combining the PROV ontology and scientific workflows for better reuse and sharing of life-science data

10h15 : Mireille Louys

Maître de conférences en informatique et traitement d'images à Telecom Physique Strasbourg, Université de Strasbourg.

Collaboratrice du Centre de Données Astronomique de Strasbourg, pour l'élaboration des modèles de données dans l'IVOA (International Virtual Observatory Alliance) pour le groupe de travail Data Model (chair 2007-2011), puis du groupe de travail Semantics pour la définition de vocabulaires pour l' interopérabilité.

Mireille Louys

Les données scientifiques distribuées en astronomie résultent d'un ensemble d'étapes de traitements qui sont décisives pour l utilisateur final qui les sélectionnera pour son étude scientifique. Dans l'infrastructure de l'observatoire virtuel, qui partage ces données au sein de la communauté astrophysique, nous avons conçu un modèle qui décrit les métadonnées nécessaires pour tracer l'histoire des données et leurs liens de génération entre différentes étapes. Les concepts de bases développés dans W3C Prov-DM sont déclinés et étendus pour notre contexte. Les services implémentant ces modèles utilisent différents formats de descriptions, soit W3C, comme PROV-N, PROV-JSON, soit spécifiques à l'observatoire virtuel comme le format tabulaire VOTable ou JSON. Le modèle sera présenté ainsi que les implémentations testées à ce jour.

10h45 : Sarah Cohen-Boulakia

Sarah Cohen-Boulakia est professeur à l'université Paris-Sud dans l'équipe Bioinformatique du Laboratoire de Recherche en Informatique.

Son domaine d'expertise porte sur la provenance dans les workflows scientifiques et plus généralement sur l'intégration et l'interrogation de données biologiques et la problématique de reproductibilité d'analyse de données scientifiques.

Sarah Cohen-Boulakia

Cette présentation dresse le bilan des travaux du groupe de travail ReproVirtuFlow du GDR MaDICS qui s'intéresse à la reproductibilité des analyses de données bioinformatiques (http://www.madics.fr/actions/actions-en-cours/reprovirtuflow/).

Nous reviendrons sur la définition de différents niveaux de reproductibilité d'une analyse (depuis la reproduction d'une analyse à l'identique jusqu'à sa réutilisation partielle). Nous introduirons les outils et familles de solutions existantes pour aider à la reproductibilité avec un focus sur les systèmes de workflows scientifiques et nous établirons un état des lieux sur les possibilités actuelles offertes, les bonnes pratiques à suivre et les opportunités existantes en recherche.

11h30 : Khalid Belhajjame

Khalid Belhajjame is an associate professor at the University Paris-Dauphine. Before moving to Paris, he has been a researcher for several years at the University of Manchester, and prior to that a Ph.D. student at the University of Grenoble.

His research interests lie in the areas of information and knowledge management. He made key contributions in the areas of pay-as-you data integration, e-Science, scientific workflow management, provenance tracking and exploitation, and semantic web services.

He has published over 60 papers in the aforementioned topics. Most of his research proposals were validated against real-world applications from the fields of astronomy, biodiversity and life sciences

He is member of the editorial board of the MethodX Elsevier journal, has participated in multiple European-, French- and UK-funded projects, and has been an active member of the W3C Provenance working group and the NSF funded DataONE working group on scientific workflows and provenance.

11h30 : Alban Gaignard

Alban Gaignard is a CNRS research engineer at l'Institut du Thorax in Nantes. He holds a Ph.D. in Computer Science from the University of Nice-Sophia Antipolis since 2013.

His research interests cover the fields of knowledge representations (semantic web, linked data) and distributed systems (workflows, large scale computing infrastructures).

He has been actively involved in a large number of projects gathering researchers and engineers from various disciplines in computer science, biology and medicine.

Alban Gaignard et Khalid Belhajjame

In this talk, we will introduce reproducibility issues in life-sciences.

The PROV W3C standard will be presented as a mean to tackle them.

We will present two research works aimed at combining provenance tracking and reasoning to address

i) multi-site workflow issues and

ii) ii) better workflow results interpretation and sharing through machine- and human-oriented summaries.

La science pour la santé
From science to health

Introduction à la provenance

Isabelle Perseil, INSERM, RDA TAB, EOSC-Life

Le W3C

Le **World Wide Web Consortium**, abrégé par le sigle **W3C**, est un organisme de standardisation à but non lucratif, fondé en octobre 1994 chargé de promouvoir la compatibilité des technologies du World Wide Web telles que HTML5, HTML, XHTML, XML, RDF, SPARQL, CSS, XSL, PNG, SVG et SOAP. Fonctionnant comme un consortium international, il regroupe au 26 février 2013, 383 entreprises partenaires

La Provenance …?

Quelques définitions usuelles

- **« La provenance est définie comme l'enregistrement des personnes, des institutions, des entités et des activités qui jouent un rôle dans la production, la modification et la mise à disposition de données ou d'autres choses. [...] Les informations de provenance font partie des métadonnées contextuelles qui peuvent elles-mêmes devenir importantes en raison de leur propre provenance. » (Groupe d'incubation sur la provenance du W3C).**

- **La provenance est fournie par les métadonnées, mais toutes les métadonnées ne concernent pas de provenance. Par exemple, le titre ou le format d'un livre constituent des métadonnées mais ne donnent pas d'informations sur sa provenance, tandis que la date de création, l'auteur, l'éditeur et les droits sur le livre donnent des informations sur sa provenance.**

In English

Provenance refers to the source of Information such as entities and processes involved in producing or delivering an artifact. (Yolanda)

Provenance is a **description** of how things came to be, and how they came to be in the state they are in today. Statements about the provenance can themselves be considered to have provenance. (Jim M)

Provenance of a resource is a record that describes entities and processes involved in producing and delivering or otherwise influencing that resource. Provenance provides a critical foundation for assessing authenticity, enabling trust, and allowing reproducibility. Provenance assertions are a form of contextual metadata and can themselves become important records with their own provenance. (W3C)

On the web, provenance would include information about the creation and publication of web resources as well as information about **access** of those resources, and **activities** related to their discussion, linking, and reuse.

Provenance is documentation of the set of **artifacts, processes, and agents** that have caused a artifact to be, and of the contexts of these entities. Provenance provides a critical foundation for assessing authenticity, enabling trust, and allowing reproducibility and assertions of provenance can themselves become important records with their own provenance. (Jim M)

Elément-clé pour décrire les évolutions d'une ressource

Les informations sur la provenance permettent de répondre aux interrogations suivantes :

- **qui est responsable de la création des données ?**
- **qui en est propriétaire ?**
- **qui a contribué à leur création ?**
- **comment ont-elles été modifiées depuis leur première version ?**
- **sont-elles affectées par d'autres données ?**
- **quels outils ont été utilisés pour générer chaque version ?**

Le modèle PROV-DM

- **PROV-DM est le modèle conceptuel de données qui sert de base pour la famille de spécifications du W3C sur la provenance.**
- **PROV-DM distingue les structures de base, formant l'essence des informations de provenance, des structures étendues pour des usages plus spécifiques de provenance.**
- **PROV-DM est organisé en six éléments, qui portent respectivement sur :**
  - **(1) les entités et activités, et le moment auquel ils ont été créés, utilisés ou achevés ;**
  - **(2) les dérivations d'entités à partir d'entités ;**
  - **(3) les agents qui exercent des responsabilités pour les entités qui ont été générées et les activités qui ont eu lieu ;**
  - **(4) la notion d'ensemble, un mécanisme nécessaire pour exprimer la provenance de la provenance ;**
  - **(5) les propriétés pour relier les entités qui font référence à la même chose ;**
  - **et, (6) les collections formant une structure logique pour leurs composantes.**

L'ontologie PROV-O du W3C

- **L'ontologie PROV (PROV-O) exprime le modèle de données PROV-DM au moyen du langage OWL (Web Ontology Language), fournit les moyens pour décrire les ontologies web structurées).**

- **Elle fournit un ensemble de classes, de propriétés et de restrictions qui peuvent servir à représenter et à échanger des informations de provenance générées dans différents systèmes et dans différents contextes.**

- **Elle peut également être spécialisée pour créer de nouvelles classes et propriétés pour modéliser les informations de provenance pour différentes applications et domaines.**

Modèle PROV-O : les Entités

- **Dans le modèle PROV, une entité est une ressource dont on veut décrire la provenance.**

- **« Une entité est un objet physique, numérique, conceptuel ou tout autre type d'objet avec des aspects déterminés ;**

- **les entités peuvent être réelles ou imaginaires. » Par exemple : un document, une partie d'un document, une idée, un article, de nouvelles, un contrat, un résultat, etc.**

Modèle PROV-O : les Activités

- **Les activités sont les processus qui ont utilisé ou généré des entités, comme par exemple : calculer un résultat, écrire un livre, faire une présentation.**

- **Les activités ne sont pas des entités. « Une activité est quelque chose qui se produit pendant une période déterminée et qui agit sur ou avec des entités ; elle peut inclure l'utilisation,**

- **la transformation, la modification, la délocalisation, ou la génération d'entités. »**

Modèle PROV-O : les Agents

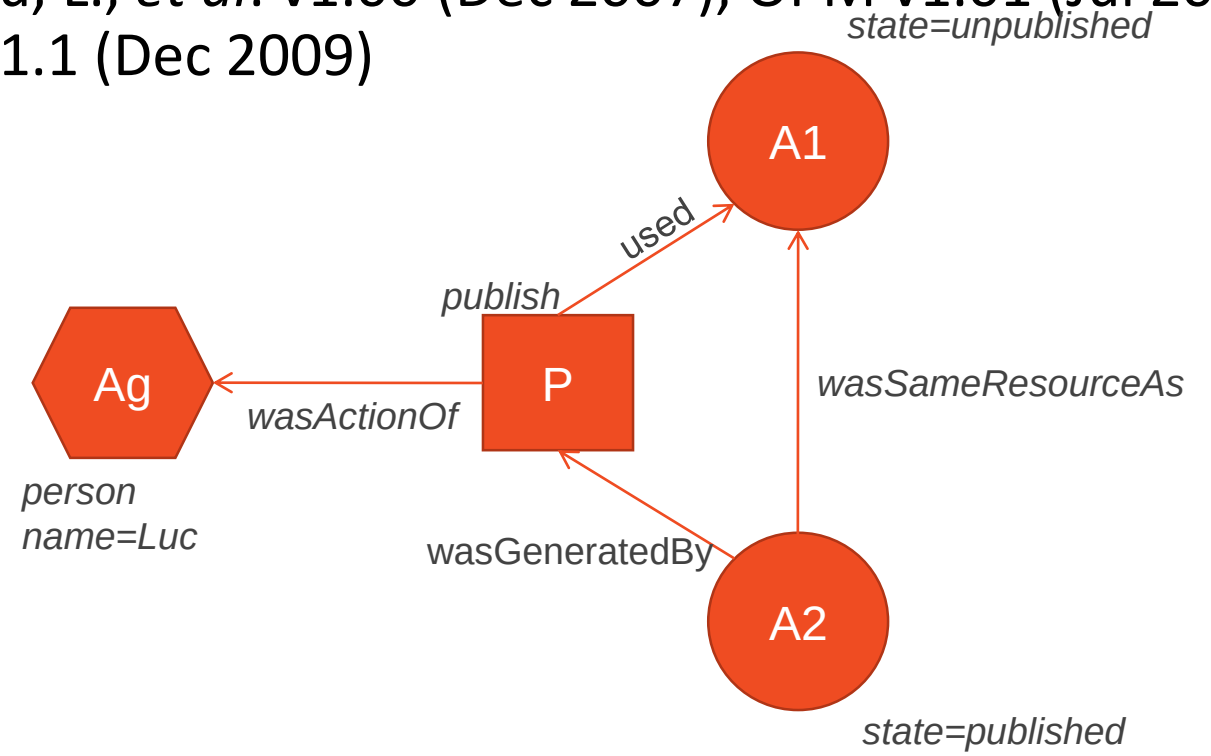
- **Les agents sont responsables des activités affectant les entités.**

- **Un agent est quelque chose qui porte une forme de responsabilité dans le déroulement d'une activité, dans l'existence d'une entité ou dans l'activité d'un autre agent.**

- **Ce peut être une personne, une composante de logiciel, un objet inanimé, une organisation, ou une autre entité.**

Open Provenance Model (OPM)

- Allows us to express all the causes of an item
- Allow for process-oriented and dataflow oriented views
- Based on a notion of annotated causality graph
- Moreau, L., *et al*. v1.00 (Dec 2007), OPM v1.01 (Jul 2008), OPM v1.1 (Dec 2009)



Nœuds: Artifacts, process, Agents

les arêtes