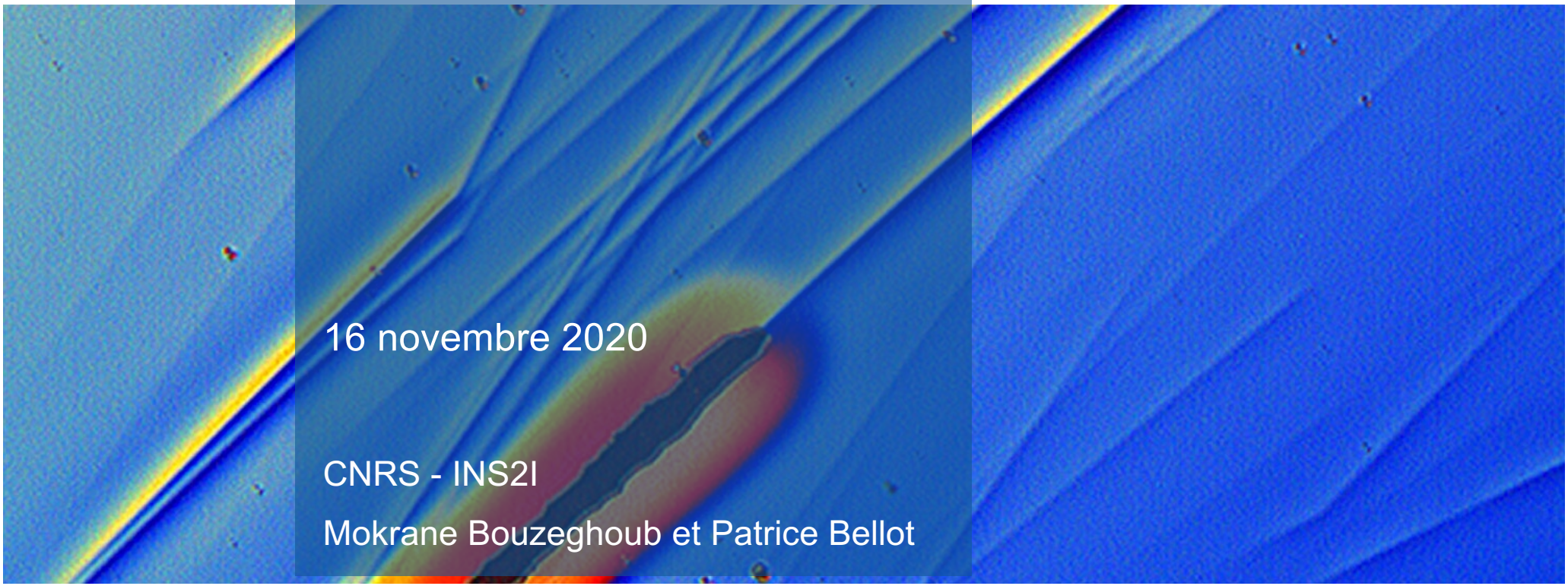
Concepts et technologies numériques pour la promotion de la science ouverte et le partage des données au CNRS

Données, logiciels, technologies



16 novembre 2020

CNRS - INS2I

Mokrane Bouzeghoub et Patrice Bellot

INS2I et Science ouverte

Introduction

- **Science ouverte vue de l'INS2I :** Une approche vertueuse et éthique de faire de la science

- **L'INS2I vue de la science ouverte:** un institut au cœur des technologies rendant possible la science ouverte

Une approche vertueuse et éthique de faire de la science

- **Accélérer la découverte scientifique et l'innovation**
  - En mettant à disposition les découvertes les plus récentes susceptibles de renforcer les théories en cours ou de bâtir de nouvelles théories.
  - En réduisant le temps d'accès au marché pour les idées ou les produits les plus avancés --> réservoir d'emplois
- **Permettre la reproductibilité et partager le contrôle des résultats**
  - Suppléer l'absence de preuve mathématique par la répétabilité des comportements des systèmes informatiques et la reproductibilité de leurs résultats
  - Transparence des algorithmes et des systèmes: Soumettre les produits à la communauté pour détecter les erreurs, les biais fonctionnels, les effets indésirables ou non éthiques

L'ouverture des données à l'INS2I - de quoi parle-t-on ?

- **A l'exception de quelques domaines, les unités de l'INS2I ne produisent pas de données de terrains (observations) ni de données dérivées (calculs) utilisables par les autres sciences (Exceptions : bio-info)**
- **Les données produites ou utilisées sont des données expérimentales construites pour valider des hypothèses ou évaluer l'efficience ou la performance des systèmes. On y introduit sciemment des biais, du bruit, des conditions aux limites, etc. pour disposer de conditions expérimentales répétables et connues**
  - Pour valider la navigation d'un véhicule autonome, apprentissage d'un algorithme
  - Pour contrôler l'évolution d'un drone dans un environnement soumis à de forts vents
  - Pour caractériser un algorithme de gestion du flux de données ou d'analyse de relations sociales sur un réseau informatique ou social etc..

L'ouverture des données à l'INS2I - pour qui et pour quoi ?

- **Les données expérimentales doivent être publiées en tant que dispositif de validation d'un logiciel donné ou de tout autre logiciel de la même classe**
- **Ces données sont issues de données de terrains sélectionnées, filtrées et calibrées de façon à répondre aux conditions de test. Elle peuvent aussi être générée automatiquement (aléatoirement) selon le type de test envisagé.**
- **Ces data sets permettent d'évaluer les comportement des logiciels en termes de:**
  - Fonctionnalités (services rendus)
  - De performance et de passage à l'échelle (temps de réponse)
  - D'efficacité (faux-positifs, faux négatifs)
  - De fiabilité (erreurs, crashs, …)
  - De responsabilité (privacy, éthique…)

Ex: : TCP data sets, Kaggle, VisualData, ...

Le logiciel : un vecteur de dissémination de la connaissance

Les unités INS2I produisent du logiciel en quantité

- **Pour des besoins génériques**
  - Compilateurs, interpréteurs de langages,
  - Gestionnaires de données,
  - Protocole de réseaux, de sécurité, et de confidentialité,
  - Langages et formats d'accès aux données ou d'échange de données
  - Gestionnaire de ressources systèmes…
- **Pour des besoins scientifiques, industriels, commerciaux ou sociétaux**
  - Contrôler un robot (véhicule, drone, humanoïde, …), surveiller un réacteur, planifier un process
  - Aider à la décision, faciliter la navigation, favoriser le commerce
  - Accéder à des connaissances, des produits culturels, des jeux…

Le logiciel à l'INS2I

- **Le logiciel est un objet de recherche pour l'INS2I**
  - Définir les langages qui le décrivent
  - Définir les méthodes d'ingénierie (conception, production, test, validation, déploiement, évolution, maintenance
  - Définir les contextes d'exécution (machine, système, réseau, …)
  - Définir les types et formats de données qu'il manipule
  - Définir les outils d'audit, de test, de validation
- **Il y a une trop grande déperdition des logiciels de recherche**
  - Coût de l'ouverture (fiabilisation et viabilisation, documentation)
  - Déficit de reconnaissances des réalisations logicielles
  - Distinction entre logiciel abouti et prototype « pour voir »
  - ➔ *La perte des logiciels peut rendre caduque un grand nombre de publications (décrivant des expériences perdues ou inaccessibles)*

Le partage de logiciel

- **Le partage d'un logiciel peut cibler deux usages distincts**
  - En tant qu'objet de recherche: Le chercheur informaticien qui peut challenger ce logiciel comme résultat de recherche, l'étendre, l'améliorer et republier à nouveau
  - En tant que nouvel instrument: pour l'utilisateur qui souhaite l'exploiter pour tester de nouvelles hypothèses ou réaliser de nouvelles expérimentations dans son domaine

- **Le logiciel libre est une des formes les plus abouties de la science ouverte**
  - Capitalisation du savoir
  - Contrôle à grande échelle
  - Évolution et déploiement rapide
  - ➔ Assez bien utilisé pour les produits génériques mais marginalement exploitée pour les produits de la recherche

L'INS2I : un institut au cœur de la science ouverte

Au delà du FAIR

FAIR : des recommandations de base pour partager les données

F : identifier universellement les objets et les décrire par des métadonnées riches

A : Les rendre accessibles via des protocoles et des interface reconnues

I : Utiliser des langages communs ou des traducteurs de modèles pour faciliter l'échange et l'intégration de données.

R : (*Comprendre Trustable ou Useful*): Décrire le contexte d'élaboration des données, origine et date de production, transformations appliquées aux données, droits d'usage…

Si les données satisfont F-A-I, la réutilisation en tant que telle ne pose pas de problème technique, les données sont réutilisables à l'infini pour peu qu'elles soient utiles.

FAIR + : Pour aller vers la science ouverte

- Adapter les recommandations FAIR au logiciel (qui est plus qu'une donnée)
- Expliciter et populariser le cycle de vie des données et des logiciels ainsi que leurs niveaux d'abstraction
- Proposer des architectures types d'intégration de données selon les objectifs de réutilisation et les pratiques des différentes communautés scientifiques
- Proposer des architectures types d'intégration de logiciels (workflows standards, architectures à base de services …)
- Populariser les outils de visualisation de données

FAIR + : pour le logiciel

- **F : Identifier universellement les logiciels en tenant compte de leur versions et de leur environnement d'exécution, mais aussi de la description multiforme du logiciel (code source, code exécutable, algorithme)**
  - plusieurs mécanismes nécessaires (SWHID, DOI...)
- **A : à la fois comme référence à citer, mais aussi comme code à invoquer dynamiquement ou interfacer avec d'autres logiciels**
  - Recherché dans des entrepôts, forges : Software Heritage, GitLab, GitHub...
  - Invoqué à travers des protocoles reconnus (API, RPC, message passing, …)
- **I (comprendre interfaçable) : avec d'autres logiciels pour échanger des données**
  - Via des architectures logicielles (Bus logiciel, web service, plateforme as a service)
  - Au sein d'un workflow scientifique (modèles de synchronisation)
- **R : sujet majeur et complexe en génie logiciel, recouvrant de multiples facettes**
  - sémantique, ressources et envt d'exec, évolution, droits, …

Pour les données : Proposer des architectures d'intégration types

- **La publication de données disparates est un pas vers le partage mais insuffisant pour la science ouverte**
  - Nécessité de capitaliser de ces données en les rapprochant et en les exploitant ensemble
- **Les communautés scientifiques ont des pratiques d'analyse de données différentes, pouvant varier d'un projet à l'autre**
  - analyse et aide à la décision, apprentissage et datamining, recherche d'information…
- **Ces besoins nécessitent des techniques d'intégration de données différentes**
  - entrepôt de données (ETL), systèmes de médiation ou de fédération de données, virtualisation et cloud… -
- **L'accès aux technologies d'intégration de données a un coût (en outillage et en expertise)**
  - Peu de systèmes open source pouvant satisfaire une large communauté

Pour le logiciel : Proposer des architectures d'intégration types

- **L'interopérabilité et l'intégration de logiciels dans des chaînes de traitements complexes fait appel à des technologies spécifiques**
  - Machine virtuelle et conteneurs
  - Architecture à base de services et services web
  - Modèle de workflows…
- **C'est une expertise multi-domaines**
  - Architectures des machines et gestion de ressources
  - Ingénierie logicielle et cycles de vie
- **Proposer des formations de haut niveau pour amplifier le mouvement d'ouverture du logiciel dans la communauté scientifique**

Conclusion

- **La science ouverte est une entreprise sociale de production du savoir**
  - Par mutualisation des connaissances
  - Partage de l'effort et des coûts
- **Ses fondements sont :**
  - Les technologies de publication et de partage des artefacts scientifiques
  - Les méta données décrivant de façon aussi riche que possibles ces artefacts pour en faciliter la compréhension et la mesure de leur utilité,
  - La transparence (plus que la confiance) des méthodes de production de ces artefacts
- **Sa popularisation a un coût**
  - en formation,
  - en soutien technique
  - et en valorisation de ses productions

Merci !

www.cnrs.fr