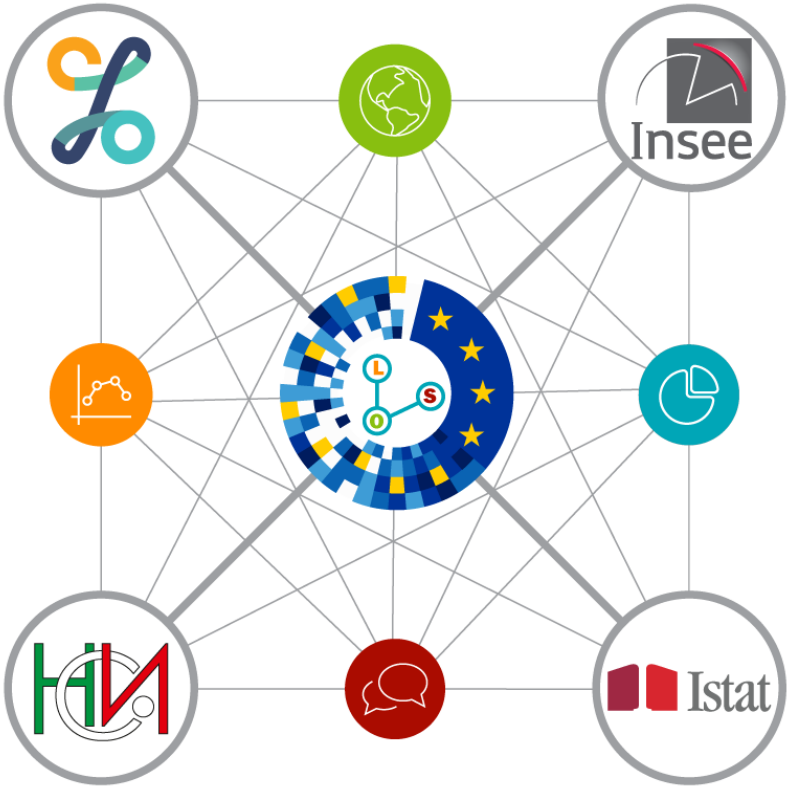


Restitution du hackathon Linked Open Statistics

Paris, septembre 2018

arnaud.degorre@insee.fr

Entre restitution… et frustration

- Exposé liminaire :
  - Le RDF, une autre façon de penser les données
  - Ouvrir, connecter, enrichir : les Linked Open Data
  - La statistique publique y échappera-t-elle ?
  - Pourquoi cet hackathon ?
- Retour d'expériences : présentation des équipes
- Mise en perspective

Web et web sémantique

Le web

- Ensemble de documents
- Briques :
  - HTTP (transport)
  - URL (adressage)
  - HTML (contenu)
- Identifier ce qui existe sur le web
- Pour les humains

Web et web sémantique

Le web

- Ensemble de documents
- Briques :
  - HTTP (transport)
  - URL (adressage)
  - HTML (contenu)
- Identifier ce qui existe sur le web
- Pour les humains

Le web sémantique

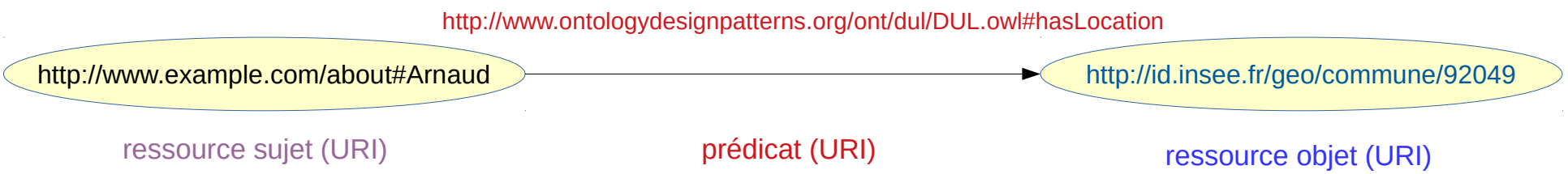
- Ensemble de connaissances
- Briques :
  - HTTP (transport)
  - URI (identification)
  - RDF (contenu)
- Identifier sur le web ce qui existe
- Pour les machines (et les humains)

RDF – Resource Description Framework

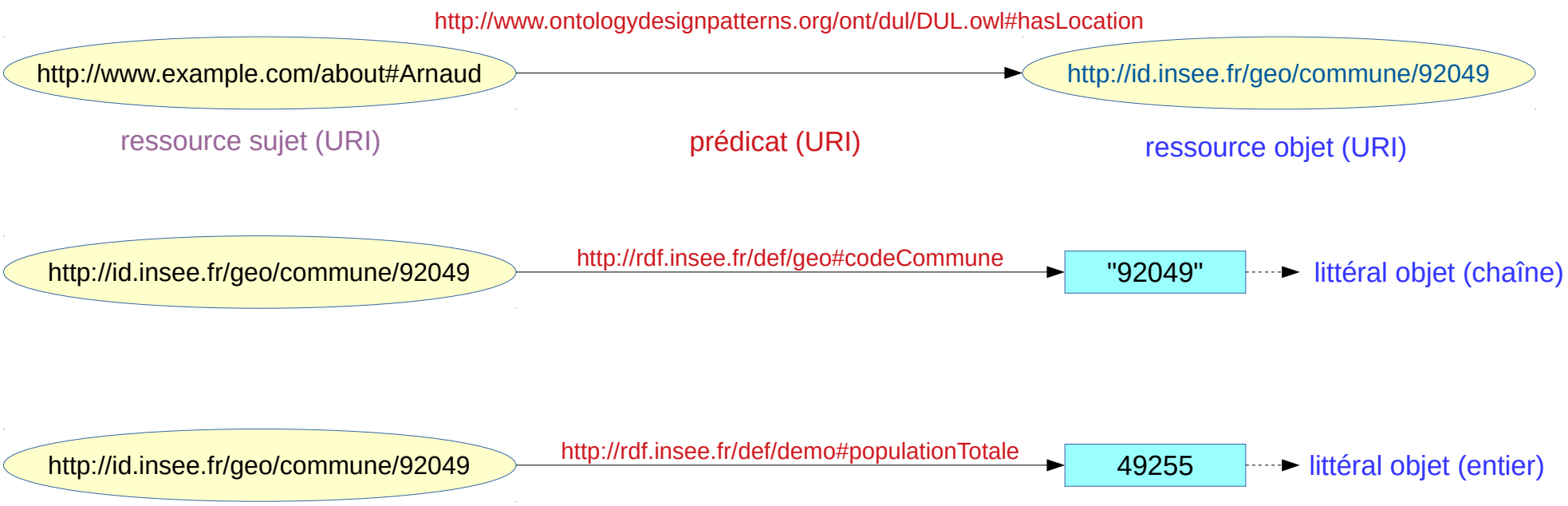
- Énoncés (« triplets ») à propos de ressources
  - (sujet) – (prédicat) – (objet)
    - Arnaud est localisé à Montrouge
    - Montrouge a pour code 92049
    - Montrouge a pour population 49255

RDF – un exemple schématisé

RDF – un exemple schématisé

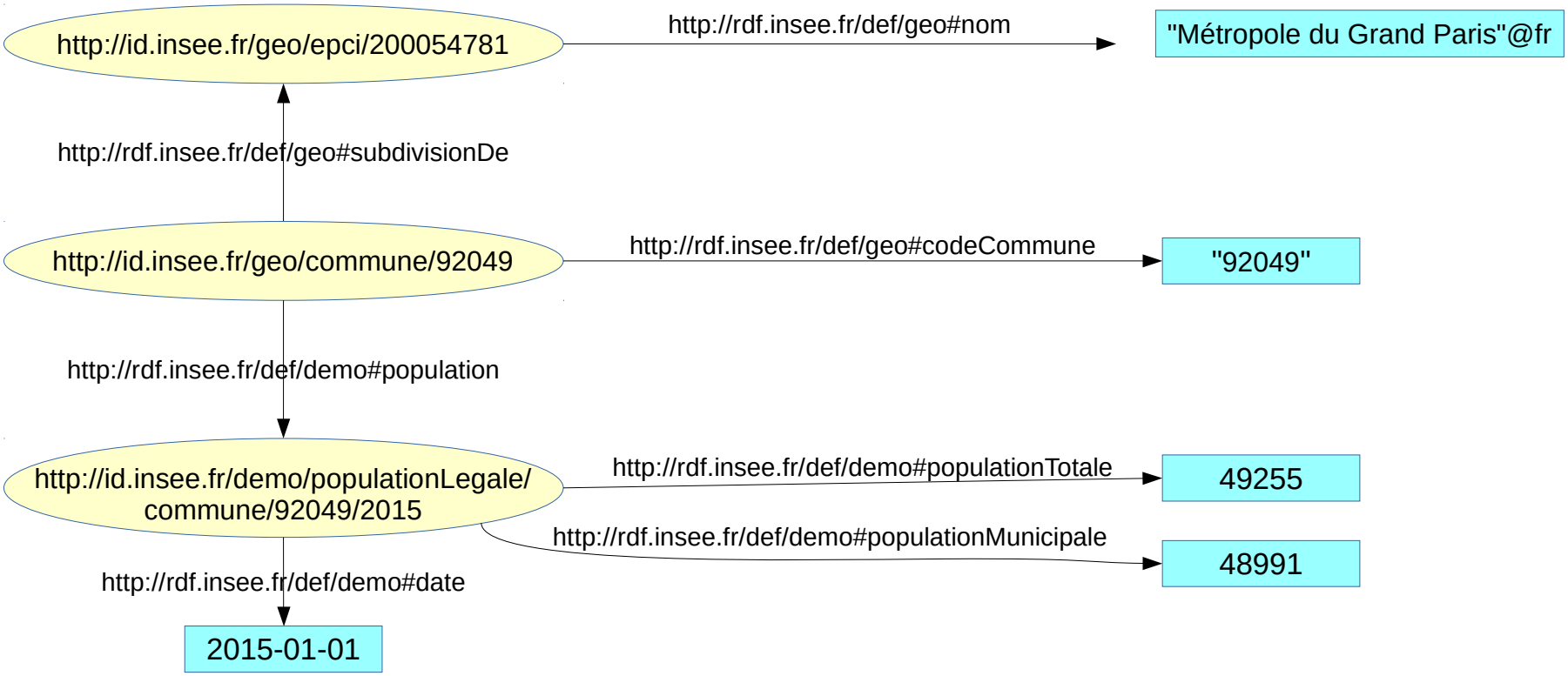


RDF – un exemple schématisé



RDF – une approche orientée « graphe »

- Les triplets se combinent en graphes



RDF – une approche orientée « graphe »

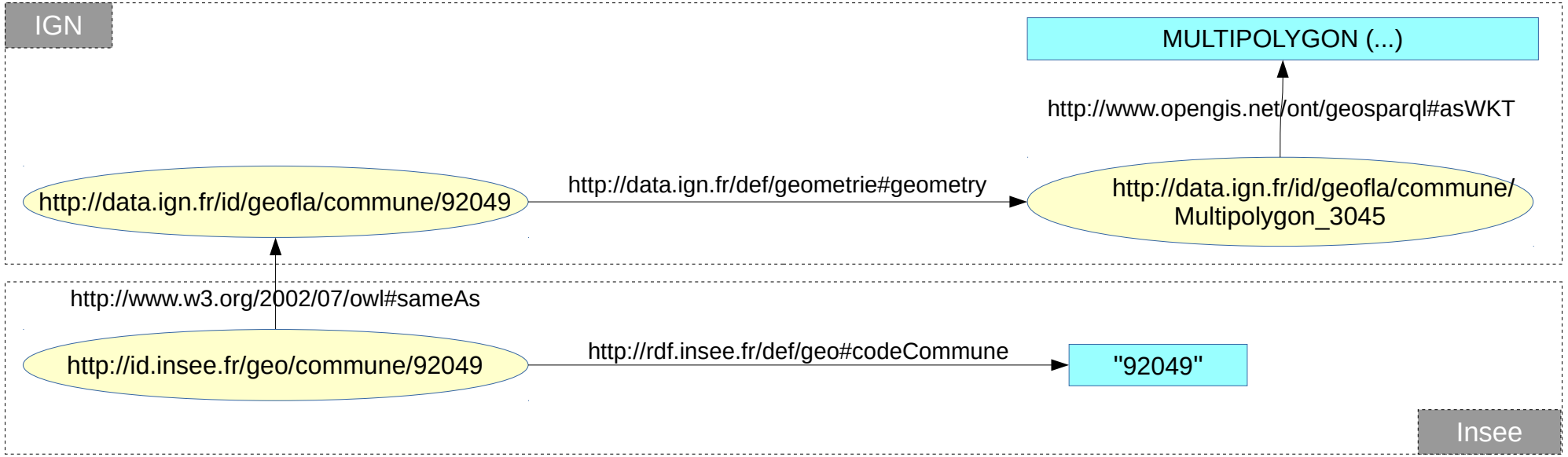
- Les triplets se combinent en graphes

RDF – une approche compositionnelle

- Les graphes se connectent entre eux

RDF – une approche compositionnelle

- Les graphes se connectent entre eux

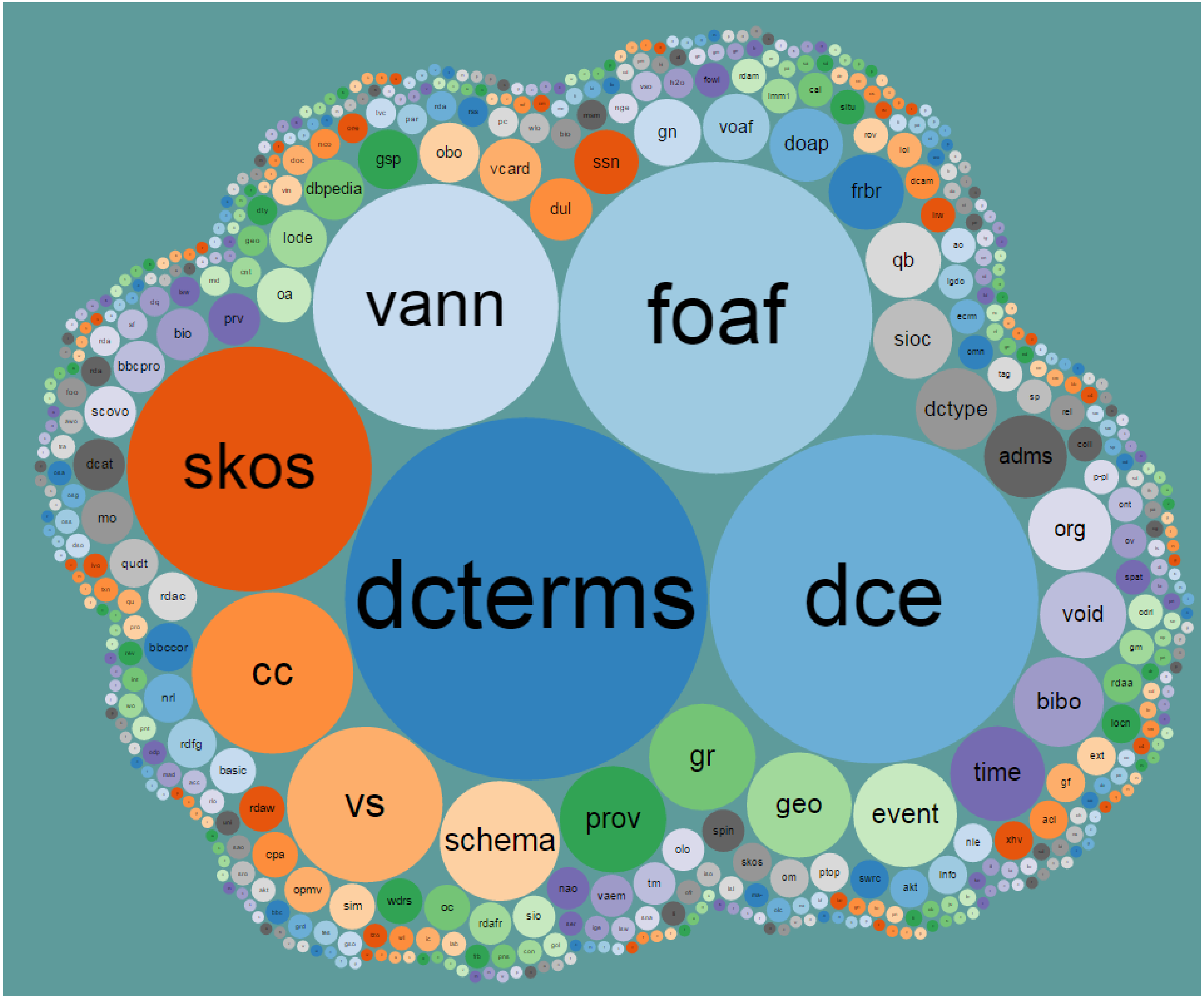


La force de la standardisation

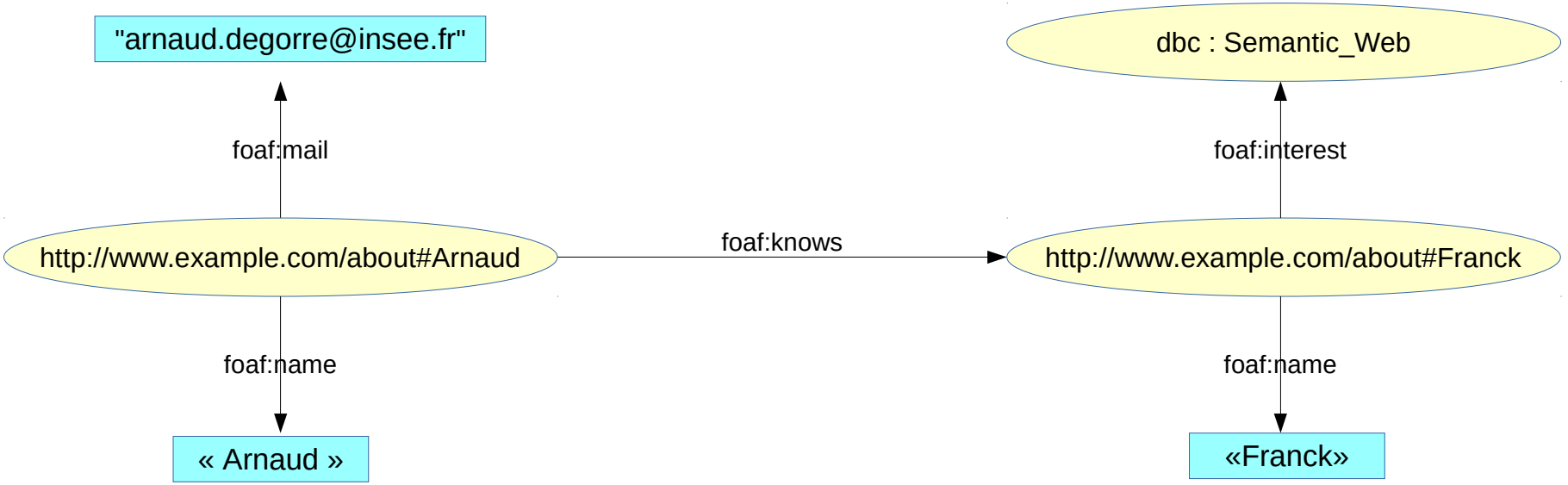
- Modèle de données universel et explicite
  - Adapté à tous types de données
  - Identification absolue des composants

- Un monde où tout est défini, dans des dictionnaires partagés
  - Une grammaire commune pour tous les utilisateurs
  - Une grammaire commune entre les silos de données
  - Des langages pour construire de façon harmonisée la représentation des connaissances et définir de nouveaux dictionnaires : RDFS, OWL

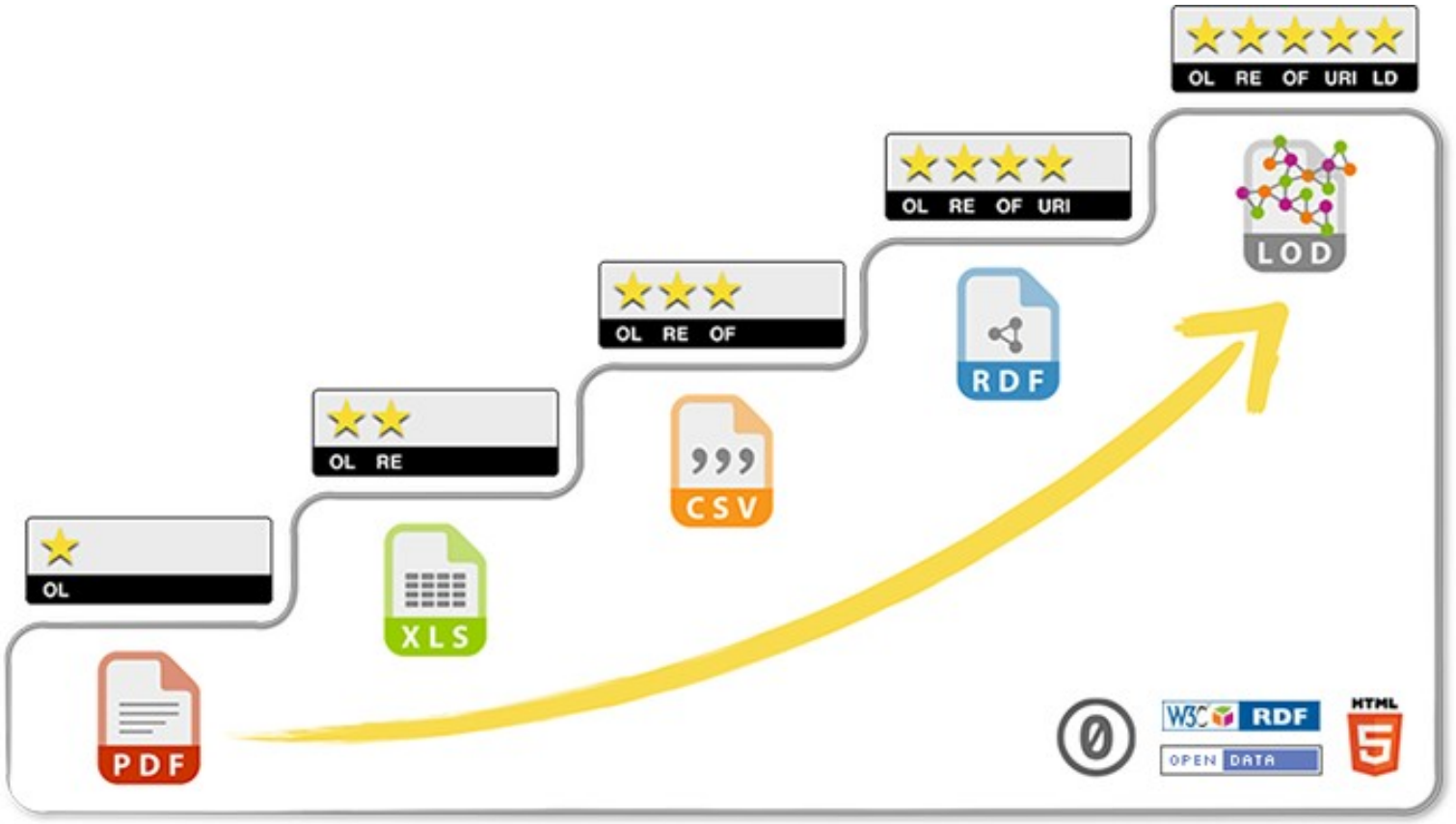
All you need is LOV



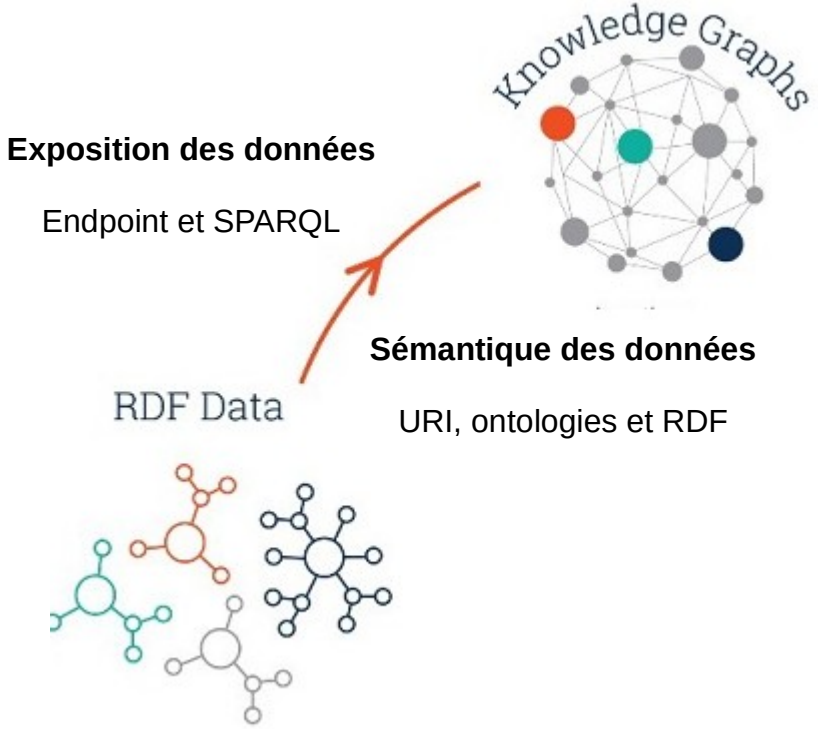
Un example d’ontologie : foaf



Tout cela pour…
exposer et connecter les données



Nuage des gisements de données



Legend

- Cross Domain
- Geography
- Government
- Life Sciences
- Linguistics
- Media
- Publications
- Social Networking
- User Generated

L'exposition des données : endpoint

- Requêter les données exposées sur http://data.bnf.fr/sparql/

```
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdarelationships: <http://rdvocab.info/RDARelationshipsWEMI/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX marcrel: <http://id.loc.gov/vocabulary/relators/>
PREFIX bnfroles: <http://data.bnf.fr/vocabulary/roles/>
SELECT DISTINCT ?docnum ?lieu ?lat ?long
WHERE
{
  ?conceptLieu foaf:focus ?lieu ;
    skos:prefLabel "Montrouge (Hauts-de-Seine, France)"@fr .
?lieu a geo:SpatialThing;
    geo:lat ?lat ;
    geo:long ?long .
?conceptLieu  skos:closeMatch ?sujet.
?edition dcterms:subject ?sujet ;
rdarelationships:expressionManifested ?exp.
?exp ?s ?p .
?edition rdarelationships:electronicReproduction ?docnum .

}
OFFSET 3
LIMIT 20
```

Pourquoi Linked Open Data et statistique publique ?

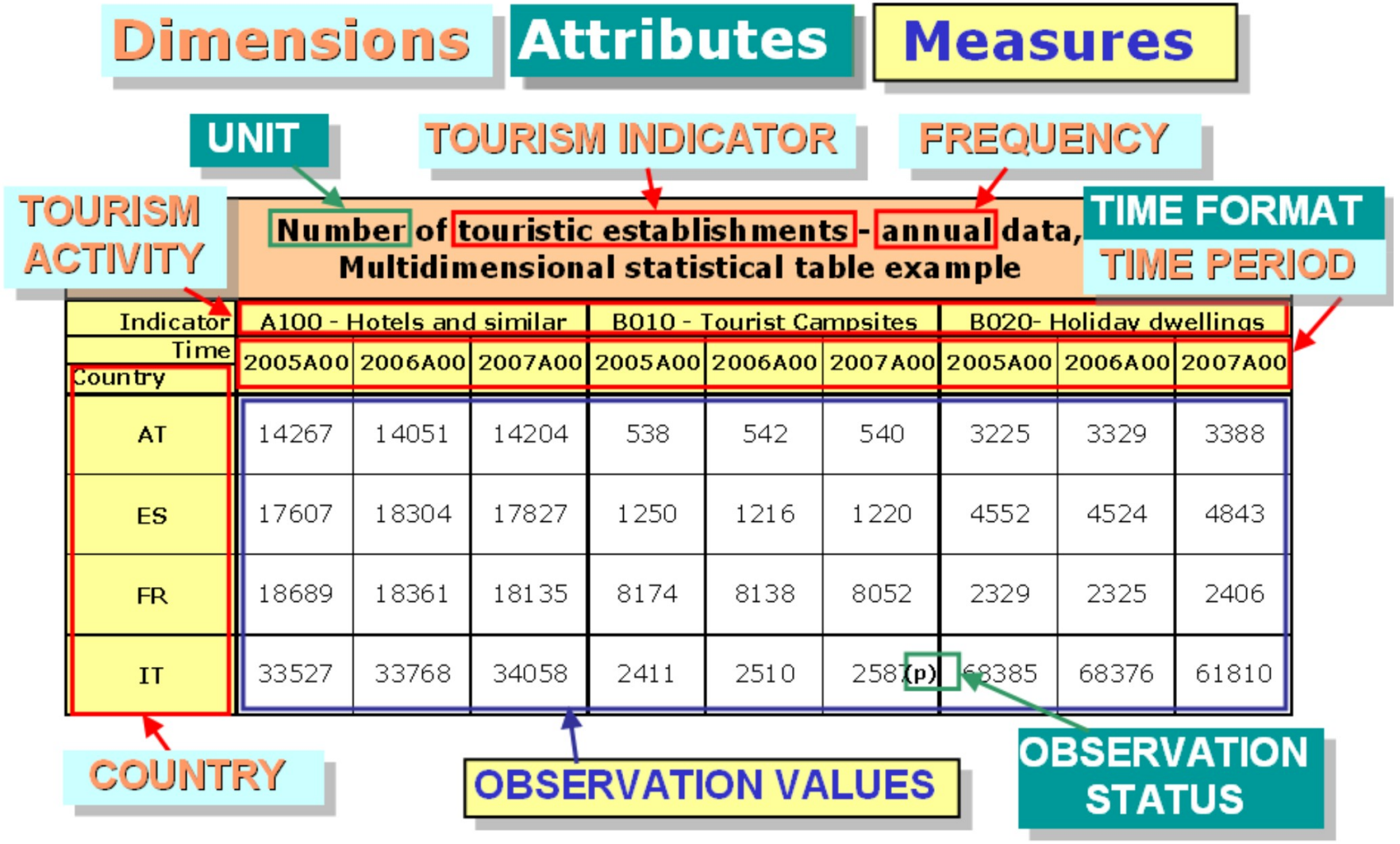
- Consolider le patrimoine statistique
  - organiser nos données dans un cadre harmonisé pour l'ensemble des INS
  - naturellement connecter nos données, via des ontologies communes
  - s'ouvrir des possibilités enrichies de requêtage
- Faire un pas de plus…
  - vers la transparence : nos données sont exposées, directement accessibles, tout est explorable
  - vers la fiabilité des usages : comparabilité construite et intégrée, documentation complète
  - vers l'intelligence numérique : exploitation par des machines, mises en relation massives de données
- Un enjeu particulier pour les registres et référentiel
  - Siren, le gisement naturel des URI pour les entreprises et établissements
  - COG, gisement naturel des URI pour les subdivisions géographiques institutionnelles
  - …

Linked open data et statistique publique

Des vocabulaires standardisés, dont :

- Données dimensionnelles : Data Cube
- Métadonnées structurelles : SKOS/XKOS
- Métadonnées descriptives : DCAT
- Métadonnées de qualité : SDMX-MM

Data Cube : http://purl.org/linked-data/cube#



Indicator	A100 - Hotels and similar			B010 - Tourist Campsites			B020- Holiday dwellings		
Time	2005A00	2006A00	2007A00	2005A00	2006A00	2007A00	2005A00	2006A00	2007A00
Country									
AT	14267	14051	14204	538	542	540	3225	3329	3388
ES	17607	18304	17827	1250	1216	1220	4552	4524	4843
FR	18689	18361	18135	8174	8138	8052	2329	2325	2406
IT	33527	33768	34058	2411	2510	2587 (p)	68385	68376	61810

l'Insee et les Linked Open Data

- Une vieillie histoire…
  - http://rdf.insee.fr depuis 2007

- … qui continue de s'écrire :
  - En transversal, RméS, « repo de métadata »
    - Définitions, nomenclatures
    - Opérations statistiques, métadonnées de qualité
  - Côté « collecte » Métallica
    - Métadonnées actives et industrialisation de la collecte
  - Côté « diffusion », Mélodi
    - Constitution de Data Cube
    - Documentation des jeux de données
    - Gestion du processus de validation et d'exposition

Eurostat et la promotion des Linked Open Data au sein des INS européens

Vision 2020

Focus sur l'utilisateur

Recherche de la qualité

Nouvelles sources de données

Efficacité des processus de production

Meilleures diffusion et communication

ESS.VIP DIGICOM

- Analyse utilisateur
- Produits innovants
- Accès aux données
  - (Linked) Open Data
  - APIs
  - Micro-données
- Communication et promotion

YOU ARE HERE

Les moyens de l'ESSNet LOS

Un partenariat avec 4 pays

- National Statistical Institute of Bulgaria (coordination)
- Insee
- Istituto Nazionale Di Statistica (ISTAT), Italie
- Central Statistics Office, Irlande

De multiples objectifs :

- Produire des cas concrets de données
- Développer des cas d'utilisation
- Tester les outils produits par le consortium
- Développer les compétences internes

Un mode d'action privilégié : les Hackathons !

« PLOSH » : Un hackathon en mode « triple »

- PLOSH, Paris LOS Hackathon

- 3 niveaux de contribution :
  - **Design :** association des producteurs à la sélection des données retenues et à leur documentation / travail préparatoire estival
  - **Produce** : conversion des données dans les formats LOD

    => *track dédiée du Hackathon, du lundi après-midi 10 au mercredi 12*
  - **Consume** : utilisation de LOD et datavisualisation

    => *track dédiée du Hackathon, du mercredi 12 au vendredi 14 matin*

Les acteurs : avant, pendant… après ?

- Le hackathon avant le hackathon
  - *Coordination : **Franck Cotton** (représentant Insee à l'ESSnet et pilote du Hackathon)*
  - *Une équipe mélant des compétences multiples, dont les acteurs de la DSI (Division Animation Conseil, Unissi, Casua, SNDI Lille…), la DMCSI (Unité Qualité, CAPR, SSPLab…), la DDAR (Division Production éditoriale)*
  - *Des producteurs de données INSEE : div. Emploi, div. Revenus et patrimoine des ménages, div. Services, div. élaboration des statistiques annuelles d'entreprise…*
  - *Des producteurs de données SSM : SDES, DEPP…*
  - *…et tous ceux qui ont contribué un peu, beaucoup, passionnément!*

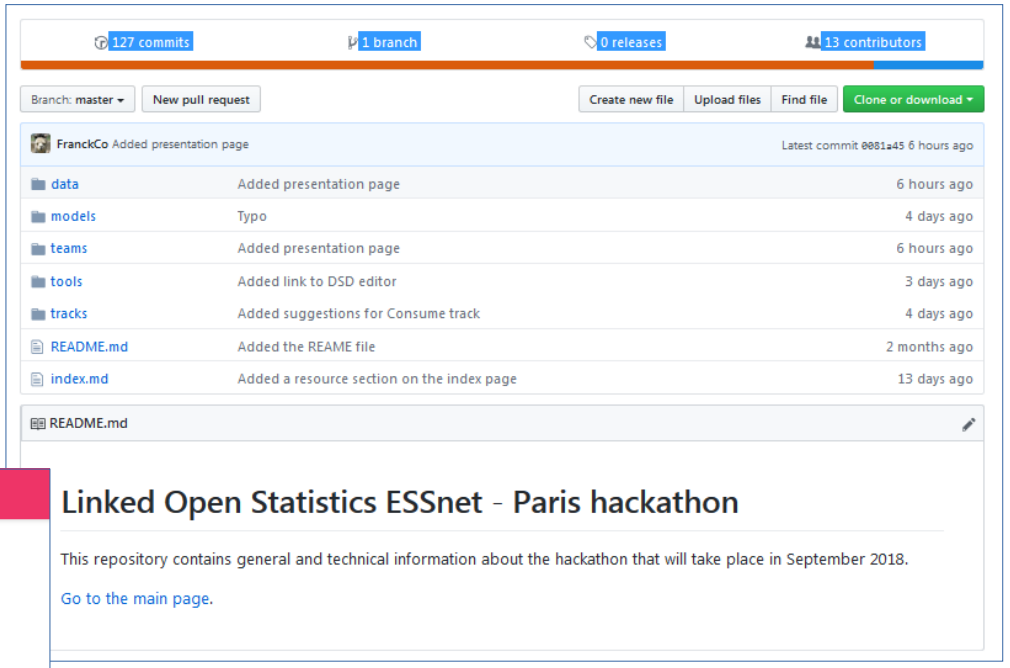
Les acteurs : avant, pendant… après ?

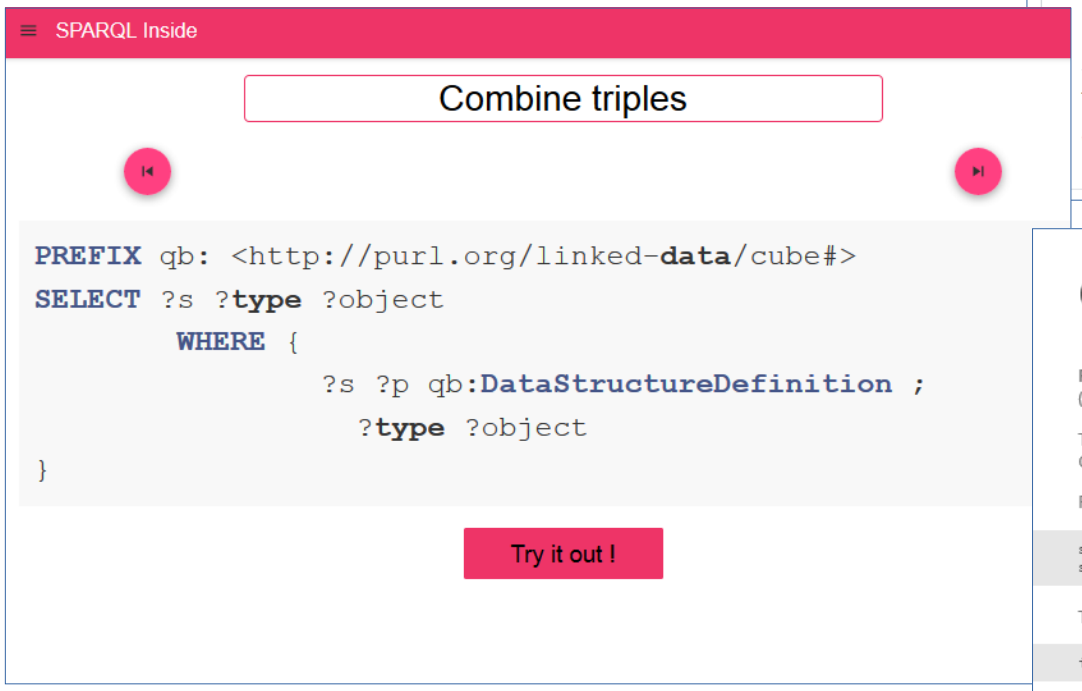
- Le hackathon pendant le hackathon
  - ***Une cinquantaine de contributeurs**, certains ayant pris part à l'ensemble des travaux. onstitution d'équipes (6 équipes par track)*
  - *Plutôt des profils « IT » sur la track « Produce », plutôt des profils « datascientists » sur la track « Consume ». De multiples langages utilisés : Java, JavaScript, Python & R*
  - *Des participants de **tous les horizons**…*
    - *À l'Insee : DG (DMCSI, DDAR, DSI…) et DR (services développement…)*
    - *En SSM : SDES, SSP, DEPP)*
    - *Mais encore : IGN, experts indépendants*

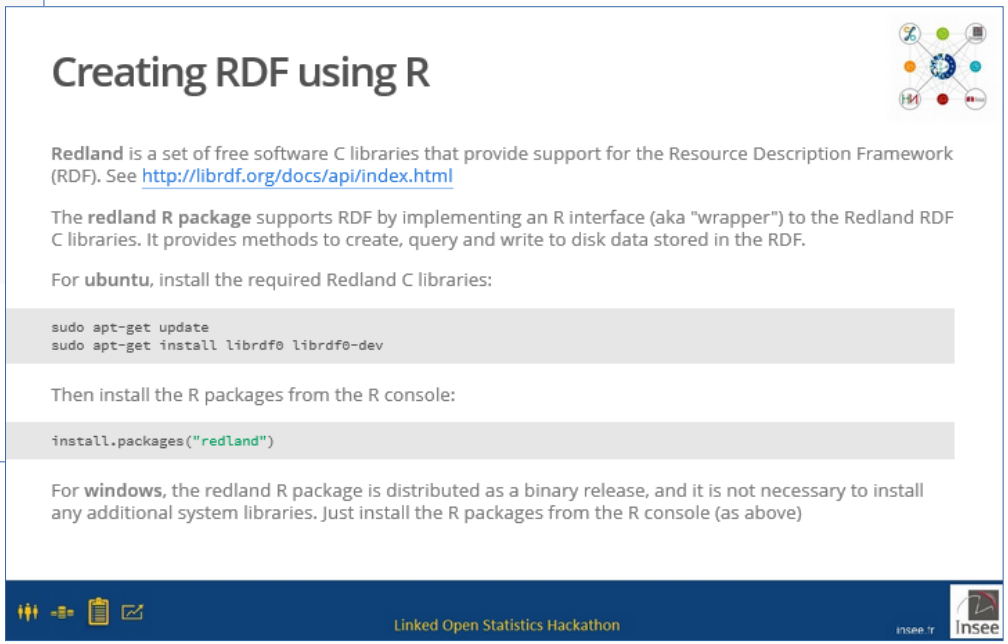
  *… y compris les **INS européens** : ISTAT, CBS, BNSI + GUS (Pologne)*

Les ressources mises à disposition…

- Un Wiki et un repo Github
- Des triple-stores
- Une plateforme Cloud







… et les ressources produites pendant le hackathon, dont des données…

Hackathon data sets

- Short introduction about data sets during hackathon is available here
- Data visualisation example here

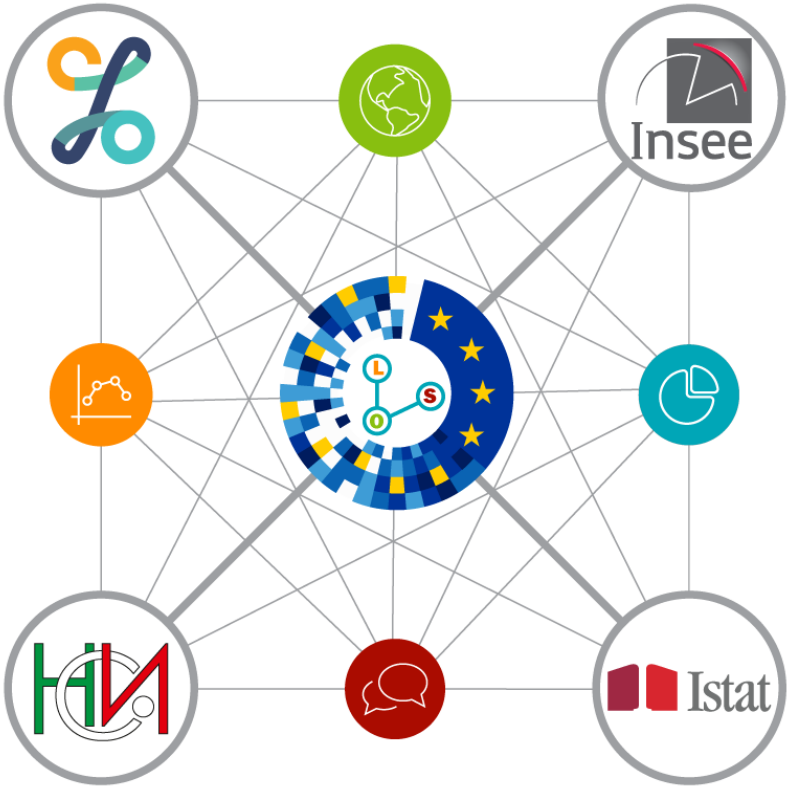
The following data sets are made available for the hackathon.

- Tourism statistics (French data set)
- Labour force survey (French data set)
- EU SILC (French data set)
- SBS (French data set)
- Census (HC55) (French data set)
- Education (French data set)
- Construction (French data set)
- EU SILC, LFS, HC55 (Bulgarian, French, Italian, Irish data sets)

Other data sources

RDF data sets or SPARQL endpoints

- NUTS (Repository: nuts)
- POP5 (Repository: pop5)
- Legal populations 2010-2015
- French geographic official code (COG)
- French classification of activities (NAF rev. 2)



… et des applications illustrant des cas d’usage (place aux démos de chaque étape)