



# Projet européen INTELCOMP

—  
**LL SANTÉ**  
**PRÉSENTATION ET POINT D'AVANCEMENT**  
**26/06/2023**



## ORDRE DU JOUR

1. RAPPEL DU CONTEXTE ET DE L'AMBITION DU PROJET
2. RETOUR SUR LES DONNÉES ET LES OUTILS DÉVELOPPÉS
3. EXPLORATIONS DES OUTILS D'IA POUR IDENTIFIER LES PUBLICATIONS ISSUES DES PROJETS FINANCÉS PAR L'INCA
4. AUTRES EXPLORATIONS POSSIBLES

Projet Intelcomp 26/06/23



## RAPPEL DU CONTEXTE ET DE L'AMBITION DU PROJET



## INTELCOMP : CONTEXTE & OBJECTIF

**PROJET EUROPÉEN SUR 3 ANS (2021-2023)**

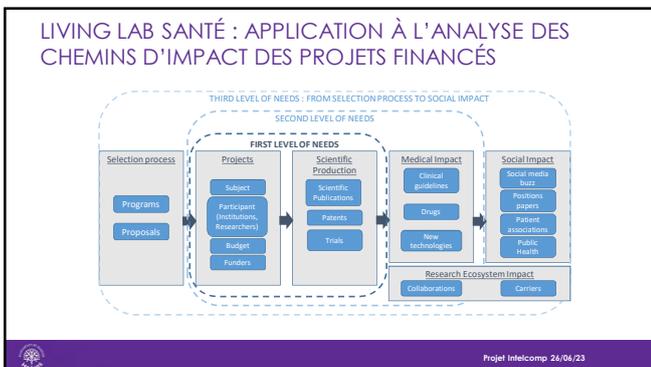
**PRINCIPAUX ACTEURS**

■ FECYT	■ OPENAIRE
■ UC3M	■ ARC
■ BSC	■ CITE
■ MINECO- SEDIA	■ HFRI
■ NTT DATA	
■ TILDE	■ OST – HCERES
■ TG	■ ZSI

**OBJECTIFS**

- Construire une plateforme de connaissance (données et outils d'IA) permettant de proposer des services d'exploration et de visualisation de données dans le domaine de la recherche et de l'innovation.
- Cible principale : **administrations publiques et acteurs des politiques de R&I** ; les services envisagés doivent aussi impliquer d'autres parties prenantes (q chercheurs, industriels, citoyens...
- Développer la plateforme en co-construction avec les utilisateurs grâce à des Living Lab : Changement Climatique, Intelligence Artificielle, Santé (piloté par l'OST)

Projet Intelcomp 26/06/23




## RETOUR SUR LES DONNÉES ET LES OUTILS DÉVELOPPÉS



## OUTILS DÉVELOPPÉS DANS LE CADRE D'INTELCOMP

### STI Viewer

- for policy makers & public administration
- to visualize (un)structured data
- to analyse datasets using NLP pipelines, machine translation, timelines

### Interactive model trainer

- for technical & advanced users
- to train their own (topic or classification) models

### STI participation portal

- for academic & industry stakeholders, citizens
- to visualize the STI panorama & its evolution, the lag between the STI outcomes and the social media impact
- to share views and feedback through Participation mailbox

### Evaluation workbench

- for public administration & funding entities
- to assist in the evaluation process of STI proposals: identify evaluators; contextualize the proposal with existing patents, publications, projects; classify proposals automatically; check similar proposal



Projet Intelcomp 26/06/23

## DONNÉES INTÉGRÉES

### FUNDER DATASETS

- Horizon Europe, Horizon 2020 and FP7 Project Portfolios (cordis.europa.eu/project) *project publication metadata crawled from the OpenAIRE API*
- HFRI Funded Projects' Portfolio (hfri.gnet.gr)
- NIH Research Portfolio (https://exporter.nih.gov) *publications, patent and clinical trials citing support from projects*

### RESEARCH OUTPUTS DATASETS

- OpenAIRE (https://graph.openaire.eu)
- Semantic Scholar (www.semanticscholar.org)
- PATSTAT (https://publication.epo.org)

### DOMAIN-SPECIFIC DATASETS: HEALTH

- Clinical Trials (ClinicalTrials.gov)
- PubMed (pubmed.ncbi.nlm.nih.gov)
- DrugBank (go.drugbank.com)

### HUMAN RESOURCE

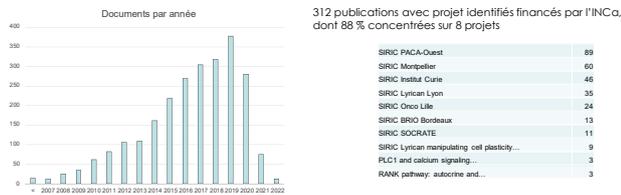
- Job Posts (tecnempleo.com)
- Euraxess Vacancies and fellowships (euraxess.ec.europa.eu)



Projet Intelcomp 26/06/23

## DONNÉES PROJETS & PUBLICATIONS INCA DANS OPENAIRE

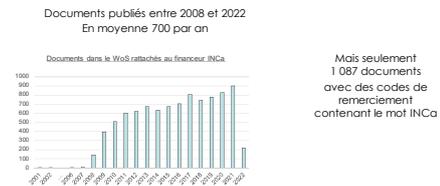
**2'206 PROJETS FINANCÉS PAR L'INCA**  
**→ 2'466 DOCUMENTS DONT 2'388 PUBLICATIONS**



Projet Intelcomp 26/06/23

## PUBLICATIONS IDENTIFIÉES DANS LE WOS

**REQUÊTE SUR LE CHAMPS « FINANCEUR »**  
**→ 9 229 DOCUMENTS ,**  
**DONT 7 482 ARTICLES ET 7 239 AVEC PMID (= DANS PUBMED)**



Projet Intelcomp 26/06/23

## PUBLICATIONS IDENTIFIÉES DANS LE WOS

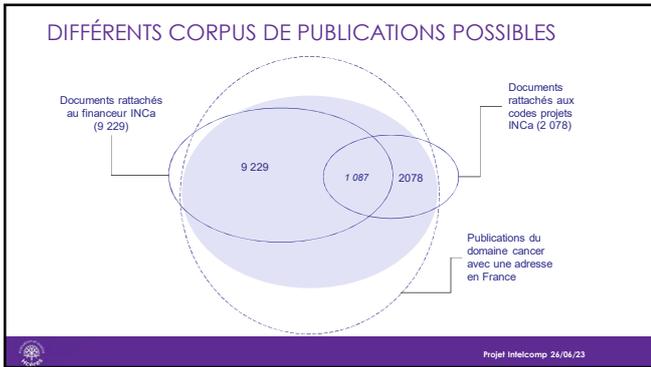
**REQUÊTE SUR LE CHAMPS « REMERCIEMENT »**  
**→ 2 078 DOCUMENTS AVEC LE CODE INCA DANS LEUR REMERCIEMENTS**  
**DONT 1 530 ARTICLES AVEC PMID (DANS PUBMED)**

Documents publiés entre 2012 et 2022  
 250 à 300 par an à partir de 2017



Projet Intelcomp 26/06/23

**EXPLORATION DES OUTILS D'IA POUR IDENTIFIER LES PUBLICATIONS ISSUES DES PROJETS FINANCÉS PAR L'INCA**



### TESTS RÉALISÉS AVEC LES OUTILS D'ANALYSE SÉMANTIQUE D'INTELCOMP

- Répartir les publications associées au financeur INCa par projet financé et AAP
- Identifier les publications potentiellement liées aux projets financés par l'INCa parmi les publications dans le domaine du cancer ayant une adresse en France

**MÉTHODE :**

Proximité sémantique entre titres/résumés des projets et titres/résumés des publications  
+  
contrôle par date  
+  
À compléter par un contrôle par organisation et nom des chercheurs financés et auteurs des publications et affiliation

Projet Intelcomp 26/06/23

### PREMIÈRES CONCLUSIONS

- CORPUS DES PUBLICATIONS AVEC RÉFÉRENCE À L'INCa FORTEMENT CONCENTRÉ SUR CERTAINS PROJETS**
- MÉTHODES SÉMANTIQUES NE PEUVENT PAS TOUT**
  - AAP orientés financement de structure → Projets retenus avec un résumé avec peu de contenu scientifique impossible à rapprocher d'un résumé de publication scientifique
  - Méthode sémantique à tester sur les AAP « thématiques » avec un contenu scientifique
  - Méthodes complémentaires à développer sur les organisations à mobiliser pour les autres AAP
  - Contrôle indispensable des résultats pour avoir une boucle de retour

Projet Intelcomp 26/06/23

### TESTS RÉALISÉS : PÉRIMÈTRE

**APPARIEMENT PROJETS - PUBLICATIONS**

- Projets des appels les plus "thématiques" / les moins "structurés"
  - 1'574, dont 1'365 avec résumé
- Publications du corpus de publications sur le cancer du WoS, restreint à 1 affiliation française au moins
  - 118'866
- Ces documents comptent 387'000 auteurs (avec erreurs de typographie) sur 1'048'000 paires document-auteur

PHRC	592
PLBIO	491
TRANSLA	209
SHESP	195
Doc	26
Doc-INESS	16
SPADOC	13
SHS	9
Doc-SHS	8
PostDoc-SHS	5
PLSHS	3
RESF-SHS	3
CHARESHS	2
RESF	1
TABMCDOC	1

Projet Intelcomp 26/06/23

### TEST RÉALISÉ : MÉTHODE

**PLONGEMENT ET SIMILARITÉ**

- Plongement des textes (titres + résumés) des projets et des documents, dans un espace à 768 dimensions (modèle specter)
- Proximité entre les plongements (similarité cosinus) entre projets et documents

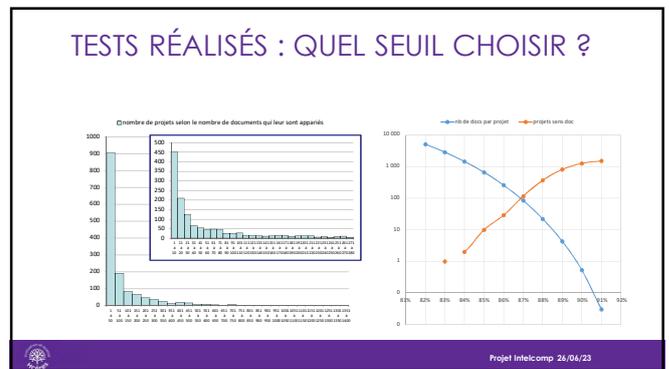
**FILTRAGE PAR DATE**

- Dates de début et de fin des projets (si absentes, reconstruites)
- Lien document-projet compatible si : début du projet ≤ document ≤ fin du projet + 1½ an

**CHOIX D'UN SEUIL DE SIMILARITÉ POUR ASSOCIER UNE PUBLICATION À UN PROJET**

- Seuil à 87 % : 36 documents en moyenne par projet ; 300 projets sans documents
- Seuil à 89 % : 0,9 documents en moyenne par projet ; 1'202 projets sans documents

Projet Intelcomp 26/06/23



## RÉSULTATS

Score	Titre	Commentaire
91,2%	Role of radiotherapy with modified Fractionation in locally advanced...	Altered Fractionated radiotherapy in the management of head and neck...
91,8%	Randomized Open Phase III Trial Testing Efficacy of Gemtuzumab Oozg...	Gemtuzumab oozgicicn for treatment of acute myeloid leukemia
91,6%	Blastic plasmacytoid dendritic cell receptors (BP2CN) in vitro and i...	CD226A-1B1 CD223 CAR T cells in blastic plasmacytoid dendritic cell ...
91,6%	NORAD01: Non-inferiority study of prospective chemotherapy without ...	NORAD01-RECCAR16 multicentre phase II non-inferiority randomized t...
91,5%	CEP-2018: A Dose-finding phase II Study of Gemtuzumab Oozgicicn in ...	Gemtuzumab oozgicicn for treatment of acute myeloid leukemia
91,5%	Clinical Benefit of Genetic Biomarkers for Guiding Treatment Decisi...	Individualised anticancer therapies: which regulatory guidelines?
91,5%	Exploitation of the anti-tumor properties of alpha-galactosylceramid...	Targeted Delivery of alpha-Galactosylceramide to CD8 alpha+ Dendri...
91,4%	Highthroughput technologies to drive metastatic breast cancer patien...	Array CGH and PKC3CA/MT3 mutations to drive patients to specific tr...
91,4%	Randomized phase III study of a treatment driven by early PET respon...	Efficacy of chemotherapy or chemo-anti-PD1 combination after failed...
91,4%	MR-FIRST 01:Improvement in the detection of aggressive prostate can...	Independent Evaluation of the Respective Predictive Values for High...

[fichier complet](#)

 Indicador de calidad Microsoft Excel

 Projet IntelComp 26/06/23



**AUTRES EXPLORATIONS POSSIBLES**



## EXPLORATIONS POSSIBLES

**STI Viewer**

- Données sur les Projets européens
- Publications issues d'OpenAire

<https://sti-viewer.intelcomp.bsc.es/indicator-report>

**Interactive model trainer**

- Corpus de publications et de brevets sur le cancer

<https://imt.intelcomp.bsc.es>

**Exploration outils sémantiques**

- Données projets INCa

 Projet IntelComp 26/06/23