NEOS - key introduction and how to participate

In this first meeting on Saturday November 6 2021 we (Pascal from France/Lyon, Rachel from Switzerland, Pol from Barcelona, Edouard from France/Paris) went into a **detailed** description of the NEOS project (what it is and what it is not - for potential complementary projects) as well as **how to concretely participate**

Sections are voluntarily short, to facilitate reading.

A. **Why NEOS ?** (Rachel I let you refine, thank you!)

PROBLEM 1: the data is here but dispersed, in diverse formats

- Ex: different hospitals in France have data represented differently. Complex to assemble
- → Need for a common language (ontology)

PROBLEM 2: much medical data, but what matters is the combination of medical data, **environment**, social and behavioral data : a “holistic” view

- → Need to assemble **environment** data (our common step here) to existing medical data, reusing existing ontologies

The latter is the goal of NEOS. In order to do so, in the following sections we describe the existing ontologies/languages and then see how to assemble environment data.

B. **Who** (@anyone: please add yourself!)

	**Pascal** is a medical doctor in Lyon (France) with a large public & private experience. His company NewClin aims at facilitating data interoperability across hospitals for example, while adding factors to health data for better decisions. In 2021 Epidemium contacted NewClin to see if he could be part of an open project, this is how this project was decided: various things here are done totally openly, contributing to the common good, while facilitating NewClin developments.

	**Rachel** is a microcellular biology researcher in Switzerland with a strong experience on DNA damage and repair. She also actively participates in Haquarium, a “biohacker” space there (a space where such projects take place). She likes to make science understood, and as such she could help on one side investigate risk factors and how they should be measured as cancer risk factors, and on the other side communicate accordingly.
	**Pol** is an architect in Barcelona and part of a team there on “data4good” - this is exactly what he looks for: using data to do good. He is also a web designer.
	**Edouard** is in data, biology and actuarial science and finance in Paris. Prior to covid he used to go to “La Paillasse”, a biohacking space. This is how he organized a similar project in 2015-16, superimposing cancer risk around the Earth with environmental, socioeconomic, food and habits aspects, to derive insights against cancer. Edouard’s key interest is anti-aging science, for a better health, notably helping solutions against cancer.
	Please add yourself !
	Please add yourself !
	Please add yourself !

C. **Leveraging from IARC monographs** (Rachel)

The IARC institution spent many human-years on listing factors that are clearly factors (group 1) or that may be (2A 2B…) factors of cancers.
Check https://monographs.iarc.who.int/list-of-classifications :

So Formaldehyde (found in some shampoos…) is dangerous. Yet...
...it depends on the dose and duration. Getting such information is key for the NEOS projet.

D. Leveraging from OSIRIS (Pol)

OSIRIS - clinical data set

OSIRIS minimum clinical data set for collection in local applications Version 1.0						
Item Group	Objective(s)	Item n°	Collection status	Item	Item definition	Expected value
1. Consent	Regulation	1.1	Mandatory	Consent date	Date of signature of informed consent form	Date
		1.2	Mandatory	Authorization for genetic analysis	Consent to proceed to personal genetic data analysis	Yes No
ification	Identification	2.1	Mandatory	Local patient identifier	Anonymized patient ID	Character string
		2.2	Mandatory	Health care center identifier	Identifier of the health care center according to the French national file	FINESS code

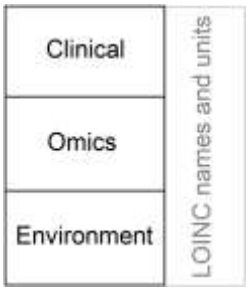
OSIRIS - omics data set

Méthode(s) et Technologie(s) de déte	Méthode(s) et technologie(s) de détection des anomalies génétiques	1.2	obligatoire (cond 1.1)	Numéro d'accession GEO	Numéro d'accession de la méthode d'analyse moléculaire dans la base de données GEO (Gene Expression Omnibus)	N° d'accession GEO (format GPLxxx)
		1.3	obligatoire (cond 1.1)	Nom de la technologie	Nom de la technologie utilisé permettant la mise en oeuvre de la méthode d'analyse.	Code OSIRIS de la technologie
		1.4	optionnel	Nom du panel de gènes	Dans le cas d'un séquençage ciblé, le nom du panel de gènes utilisé. Il peut s'agir du nom d'un panel « maison» ou commercial.	OSIRIS:O10-0 : Not applicable OSIRIS:O10-1 : Ion AmpliSeq Cancer Hotspot Panel v2 OSIRIS:O10-2 : Ion AmpliSeq Colon and

GitHub's OSIRIS

E. What we aim at

Here is a view of what we could try to reach:



As schematized by this image, the goal is to start with the Clinical and Omics part of OSIRIS (therefore it is very important to be initiated with the previous section) and to add the Environment part.

PS: On the right of this image, “LOINC” (loinc.org) is the world's most widely used terminology standard for health measurements, observations, and documents. Concretely, it is a language to name our variables and choose units (ex: mg or g). The use of LOINC is optional in this project (in order to focus and deliver), but it is good to mention it in case it is easy to be LOINC-compliant while doing the project

F. **How / splitting tasks**

Under construction:

1) Standardisation (structuring, terminology, interoperability) of environmental cancer risk factors

- 1a. [one to a few hours] Build googledoc table (sort of excel file) with the osiris table (see above) and the environment agents from IRC monographs (see above)
- 1b. [2x[1] one to a few hours] On the right of the googletable, make the link with LOINC

2) The identification of open sources for the corresponding data in France

- 2a. [half a day?] On the right of the googletable, add elements from MyDataBalls
- 2b. [a day?] List sources from data.gouv.fr that may correspond to what is needed
- 2c. [after 2b] Try to make the match with the googletable: add a column that suggest matches
- 2d. [after 2b] Go through sources of https://github.com/Epidemium/GeoStatus/raw/main/interesting_data_sources/EpidemiumArcheology.docx

3) The definition of variables to be part of the final set of data, in terms of measure, granularity, exposure level.

- 3a. [1 hour] Make five-column googledoc table with the list of variables (the one from IARC monographs), the names of contributors (people will put their names), the date of start, the advancement
- 3b. [long - transversal - after 3a] Follow the table 3a , a=contact people when not advancing to see what is ongoing, finding someone to replace, etc.
- 3c. MANY PERSONS [1h for many persons or many time] Pick an agent and search the IARC monographs and if not then Pubmed or Google scholar for an article such as ???

4) Cherry on the cake: do CancerMortality = f(environment) at department level, based on the collected data.

[1] Investigate the presence of the IRC monograph agents in LOINC.org (download tables). Then , once 1a is finalized, complete the googledoc table

- 4a. [2 days ?] Prepare (in R or python - please indicate for a better coordination) cancer mortality rates by age at **department** level: https://www.santepubliquefrance.fr/maladies-et-traumatismes/cancers
- 4b. [2 days ?] Prepare a randomForest model in R or python (same as 4a) (for example; one that can use many explanatory variables) with any variable of choice at department and year granularity and study its behavior with a toy model. Ex: to what extent, if many risk factors are purely random numbers and just a few variables make sense such as unemployment and social status you can still regress CancerMortality = f(environment) and detect that unemployment and social status are the ones that make sense
- 4c. [the greatest part] Apply it to the environmental data as they come in the project