

## IR production, traitement et analyses de données. CNRS BAP D, 2023

### Diap 1

Mesdames et messieurs les membres du jury, bonjour, je m'appelle Thomas Huet. Je présenterai mon cursus, mon expérience dans la science ouverte et en ingénierie de la recherche, et je conclurai par mes motivations à rejoindre les laboratoires ArchAm et Trajectoires comme IR.

### Diap 2

Préalablement à mon doctorat:

- 2000-2: Maîtrise d'archéologie sur l'agriculture inca à Paris 4.
- 2003-4: fouilles au Pérou avec l'université La Catolica. Je **passerai** un brevet de topographie à l'université d'Ingénierie de Lima
- 2004-5: De retour en France, je suis un DUT génie informatique au CNAM
- 2005-12: Master 2 et thèse à l'université Nice Sophia-Antipolis, au CEPAM-CNRS, sur l'étude des gravures rupestres du mont Bego

Entre la fin de mon doctorat en 2012 et aujourd'hui, j'ai accumulé 7 ans de contrats dans la recherche et l'ingénierie de la recherche en archéologie dans huit laboratoires et trois différents pays. J'ai suivi de nombreuses formations, notamment en statistiques et en traitement d'image (2D, 3D) et j'ai acquis une longue pratique de la programmation informatique et des SIG/base de données.

Durant mon contrat avec Archaïos (2019-20) je me suis familiarisé avec la BD EAMENA. C'est entre autres grâce à cette expérience que j'ai été recruté par l'université d'Oxford en 2021 pour devenir responsable de cette BD, avec un contrat courant jusqu'en 2024.

### Diap 3

EAMENA est une BD, *open access*, sur l'archéologie et le patrimoine en danger qui regroupe, entre autres informations, 180,000 sites répartis de la Mauritanie à l'Iran. C'est une instance d'Arches, un système d'information orienté websig pour l'inventaire et la gestion du patrimoine culturel immobilier. Arches intègre une pile de logiciels libres qui vont d'une DB Postgres à une interface spatiale Mapbox. Elle s'appuie sur de nombreux standards du web sémantique allant du modèle CIDOC-CRM jusqu'au 3-IF.

Mon travail comme responsable de la BD EAMENA, implique des tâches d'administrateur système, d'administrateur BD, de gestionnaire BD et d'analyste de données. Je participe directement à la formation des utilisateurs et des gestionnaires BD pour les clones nationaux, à la rédaction des cahiers des charges, à la politique de développement de la BD: notamment en ce qui concerne son internationalisation, son accès, sa définition et le signalement de la propriété intellectuelle des jeux de données qu'elle conserve, et à sa maintenance sur le long terme.

### Diap 4

J'ai inscrit mes activités sous le signe de problématiques variées et dans la recherche de solutions de mise en œuvre. J'ai diffusé les résultats de ces travaux dans des publications

scientifiques, des colloques, et en participant à des programmes et projets internationaux. J'ai aussi participé à l'évaluation d'articles et de logiciels orientés recherche, et science ouverte, à la transmission des connaissances par des enseignements et des encadrements universitaires.

Ces différentes collaborations m'ont permis de tisser un réseau national et international. Elles m'ont appris à comprendre l'existant, à analyser les besoins, à être force de proposition, à établir des protocoles, à gérer des projets dans un temps donné. Je sais m'adapter, dialoguer avec les chercheurs, reformuler leur besoins pour leur proposer **des solutions informatiques adaptées.**

## Diap 5

J'en viens maintenant à mon expérience en **production, traitement et analyse de données.**

1, ces données sont de différents types: des données alphanumériques, des images, des données spatiales, des références bibliographiques. De différents formats, elles viennent avec des métadonnées, des annotations, des incertitudes, des licences et un versionnage qu'il s'agira de traiter.

2, aujourd'hui, avec la massification et l'ouverture des données, il devient probable qu'un seul projet ne puisse exploiter l'ensemble du potentiel des données qu'il a généré.

3, pour partager ces données, il s'agit de passer de données structurées, à des données alignées, puis à des données standardisées pour permettre leur réutilisation et produire de la sérendipité.

4, suivre les principes FAIR doit permettre la réutilisation des données qui est l'objectif ultime de la science ouverte, et la principale motivation du plan de gestion des données.

## Diap 6

J'ai plusieurs fois mis en œuvre tout ou partie des principes de la science ouverte: données ouvertes, code source ouvert et accès libre.

Par exemple à travers la publication du package R, *iconr*, sur le dépôt officiel du CRAN. Ou encore en publiant des catalogues révisés de dates radiocarbone, NeoNet, sous la forme de fichiers avec un DOI, d'un article de données en libre accès, en créant une application web Shiny lisant ces données, en publiant le code source de cette application, et finalement en créant un connecteur permettant de lire ces données dans un agrégateur de dates radiocarbone.

## Diap 7

Pour la **production et acquisition des données**: je maîtrise de nombreuses techniques liées aux images aussi bien en 2D qu'en 3D, et à différentes échelles: microscopie, macrophotographie, photogrammétrie, photogrammétrie aérienne, GPS différentiels, stations totales, et télédétection. J'ai par ailleurs diverses expériences de fouilles ou de prospection (dans 5 pays), et participé à l'étude ou la révision de collections archéologiques. Je pourrais donc être utile aussi bien sur un terrain, dans les réserves d'un musée, qu'en laboratoire.

## Diap 8

Le **traitement de données** est l'ensemble des processus qui suivent l'acquisition des données brutes et précède leur analyse: cela recouvre la conception et le développement de bases de données.

A titre d'exemple, en 2016, à l'échelle de la fouille programmée d'une grotte, j'ai développé un système d'information, intégrant une BD Postgres, des relevés photogrammétriques géopositionnés avec une station totale, des sorties SIG et des visualisations 3D avec CloudCompare et R.

Plus récemment, en 2022, dans le cadre d'EAMENA et à l'échelle d'une région, j'ai automatisé la création de cartes et de profils de chemins caravaniers avec le package eamenaR et un requêtage à la volée des altitudes des caravansérails via des API REST.

## Diap 9

Mes apports à l'**analyse des données** couvrent différents domaines, toujours en lien avec l'archéologie. J'ai ainsi réalisé une analyse forme pour des figurines en silex, dans un contexte du Néolithique ancien du Proche-Orient. Ou encore en participant à la modélisation l'évolution de la triade domestique dans un contexte de la longue durée (fin du Bronze jusqu'à l'Antiquité tardive) dans le NE de la péninsule ibérique à travers des statistiques descriptives, exploratoires et confirmatoires.

## Diap 10

J'ai aussi participé à la modélisation des échanges et réseaux de circulation dans les Alpes centrales durant le premier âge du Fer, avec une approche mêlant plus courts chemins et tracés de graphes. Ou encore en créant des routines d'analyse et de modélisation de la diffusion du premier néolithique en Méditerranée centrale et nord-occidentale. Les données et le code source de l'application websig interactive Shiny que j'ai développé, ont été déposés sur Nakala.

Je conclurai par mes motivations à rejoindre les laboratoires ArchAm et Trajectoires comme IR en production, traitement et analyses de données.

## Diap 11

A la fois archéologue et informaticien, je maîtrise le cycle de vie des données archéologiques, différents outils d'Huma-Num, les principes FAIR, des langages de programmation, bases de données, l'architecture client-serveur, le versionnage des données et du code informatique, la production logicielle, les technologies du web, ainsi que de nombreux logiciels libres.

Ma principale motivation vient de ce que, selon moi, les comparaisons et la mesure des différences sont à la base des SHS, le principal défi de pouvoir rendre ces comparaisons reproductibles et transposables, entre différentes aires chrono-culturelles et pour différents types assemblages archéologiques (*what*). Dans un contexte de science ouverte, ces défis sont d'autant plus intéressants à relever.

La création d'un poste d'IR mutualisé entre deux laboratoires travaillant sur l'archéologie pré-protohistoriques des Amériques et d'Europe implique une convergence dans la structuration, l'alignement, la standardisation, et les routines d'analyses des données produites entre les labos.

Je vous remercie de votre attention.