HyperThésau Bibracte Numérique

Archéologie :
le trésor est noyé
dans la brume des données,
on cherche un phare…
ou une boussole

BIBRACTE

Archéorient
Environnements et sociétés de l'Orient ancien

Présentation : Jean Pierre GIRARD (Archéorient)
Sébastien DUROST (Bibracte EPCC)

Bibracte Numérique

BIBRACTE

Historique :

Premières fouilles sur le mont Beuvray dès 1864, jusqu'en 1907

Analyse des découvertes aux origines de la pratique archéologique moderne

Reprise de fouilles annuelles depuis 1984, par des équipes internationales

Carte d'identité scientifique :

Bibracte-mont Beuvray (Morvan, Bourgogne) :

– site archéologique majeur (Grand site de France)

– musée de France (42 000 visiteurs/an)

– Centre de conservation et d'étude pour la Région Bourgogne

Données, publications et archives :

Système d'enregistrement de données centralisé depuis 1993

Service de publication scientifique intégré

Archives : 150 ans de publications scientifiques en bibliothèque, numérisées et océrisées

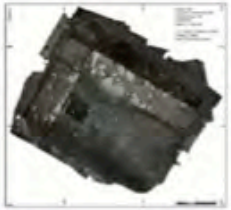
Toute la chaîne opératoire de l'archéologie – « cycle de vie » du terrain au grand public

Bibracte Numérique

BIBRACTE








Bibracte :

hétérogénéité des données

HyperThésau

Hyper thesaurus et lacs de données :
fouiller la ville et ses archives archéologiques

Archéorient
Environnements et sociétés de l'Orient ancien

Disciplines :
[1] Archéologie,
[2] Informatique, [3] Sciences de l'information
[4] Écologie-Sciences participatives

Laboratoires :
[1] Archéorient (FR MOM), ArAr (FR MOM), Universitat Autònoma de Barcelona
[2] ERIC (MSH Lyon), [3] FR MOM, Persée-ENS
[4] CESCO (Museum national d'Histoire naturelle)

Partenaires :
Bibracte EPCC, Service archéologique de la Ville de Lyon, Archeodunum SAS,
site-musée d'Ullastret (Museo dArqueologia de Catalunya)

Questions et objectifs de recherche :
Enrichir et décrire des jeux de données archéologiques et les réunir dans un lac de données pour les rendre trouvables, accessibles, interopérables et directement réutilisables (FAIR)

HyperThésau

Hyper thesaurus et lacs de données :
fouiller la ville et ses archives archéologiques

Archéorient
Environnements et sociétés de l'Orient ancien

Questions et objectifs de recherche

1- Création et déploiement de micro-thesaurus « archéologie-métier »

2- Implémentation d'un « lac de données » archéologiques semi-structurées mais à la sémantique floue et grevées d'incertitudes multiples (chronologie, identification…)

3- Fouille du « lac » et diffusion des données (« cycle de vie ») en accès ouvert, pour leur ré-exploitation et leur médiation, du laboratoire au musée

Quatre objectifs distincts font l'objet des *work packages*

1- Éléments d'un thesaurus-métier pour l'interopérabilité des données archéologiques

2- Modèle de métadonnées d'interrogation (normes du web sémantique)

3- Architecture de stockage et de fouille de données hétérogènes (le « lac »)

+

4- Une expérience d'enrichissement participatif de données d'archives archéologiques

Porteurs de projet : Marie-Odile ROUSSET (Archéorient)
Jérôme DARMONT (ERIC)

HyperThésau Bibracte Numérique

État des lieux : circulation de la connaissance archéologique (fouilles préventives et programmées, en France)

1- **Rapport** final d'opération (RFO) → évaluation par l'État – seule obligation

2- **Publications** sur papier et support numérique ← évaluation par les pairs

Les données à la source de la recherche **ne sont pas diffusées**, elles sont partagées « de la main à la main » dans le réseau scientifique (ouverture réelle mais arbitraire).

3- **Médiation** vers le grand public – principalement muséographique

= exploitation dérivée des publications scientifiques

HyperThésau Bibracte Numérique

« Vision » et objectifs communs

La « **donnée** » au centre de la démarche (avec sa description)

Partage avec tous les publics :
– chercheur
– visiteur du site
– visiteur du musée
– sur la Toile

Principes **FAIR** – Plan national pour la science ouverte
(accès ouvert aux publications et aux données, dynamique internationale, normalisation)

HyperThésau Bibracte Numérique

« Vision » et objectifs communs

« Cycle de vie » éditorial :
– Données de la recherche : sources « primaires » pour la recherche et la validation de ses résultats ; base de toutes les études, analyses et « inférences » ultérieures
– Articles de données (*data papers*) : description fine d'un jeu de données en vue de leur compréhension-appropriation et de leur réutilisation
– Publications : études et analyses spécialisées soumises à validation par les pairs

« Cycle de vie » élargi : Enrichissement(s) au fil de l'eau (« curation ») – Données enrichies/transformées par les analyses spécialisées et les études thématiques – Médiation et muséologie/muséographie – Conservation à moyen et long terme

Archéologie : de quoi parle-t-on ?

HyperThésau Bibracte Numérique

Dans la brume des « données brutes » en archéologie

1- La matière scientifique, ce sont des « traces de traces » (analogiques, puis numériques)

2- La « donnée » de terrain n'existe pas : elles est construite et « structurée » par nature

3- Toutes les tentatives d'unification technique de la représentation des données se sont brisées sur un double mur (Lukas 2018) :
– l'évolution des technologies et de leurs formats explicites/formalisations implicites
– l'évolution continue des techniques et des connaissances intrinsèque à la discipline

La **diversité des protocoles** reflète la **diversité des points de vue**

La **diversité des vocabulaires** reflète la **diversité des terrains, des usages, des traditions**

→ Au pire : imprécision – au mieux : **incertitude**.

HyperThésau Bibracte Numérique

Contexte : la nébuleuse (très) évolutive des types de données

1- Archives XIX^e s. → support papier

– publications savantes

– plans, croquis, dessins

– carnets de fouille manuscrits

Nécessité d'une **rétro-documentation** (numérisation, océrisation, métadonnées)

2- Fouilles modernes → support numérique/numérisé

– pratiques de terrain : invasives et non-invasives ; gestion de la stratigraphie ; nommage et regroupement des faits archéologiques

– systèmes d'enregistrement : aucune normalisation ; évolutivité et temporalité des données (« cycle de vie scientifique ») généralement non gérées

– synthèses et études thématiques (monnaies, céramique, etc.)

HyperThésau Bibracte Numérique

Contexte : la nébuleuse *des* archéologies

Les partenaires de BibNum et HyperThésau reflètent la diversité de la discipline :

– archéologie programmée et préventive, publique et privée (ville de Lyon, Archéodunum)

– archéologie en France et à l'étranger (Morvan, Catalogne, Proche-Orient)

– diachronie : âges du Bronze et du Fer, Antiquité, Moyen Âge

– focalisations thématiques : monnaies, objets, archéologie du bâti, etc.

– imagerie (géophysique, etc.) et données spatialisées

et de son rapport à ses publics :

– musées de France… et d'ailleurs : Bibracte, Ullastret

– plateformes de publications numériques et « *triple store* » : Persée

– plateforme de « science participative » → multitude des amateurs *engagés* : 65 MO

HyperThésau Bibracte Numérique

Différentes approches « post-fouille » de la complexité

1- Ne rien faire : données isolées, analyses solitaires, publication sans jeu de données associé

2- Mettre à disposition et signaler, sur un entrepôt public, les rapports, analyses et études spécialisées, documents graphiques et publications associées

3- Publier sur un site web et référencer (SEO) les rapports, analyses, documentation, etc.

4- Structurer en TEI une publication et les « données d'autorité » qu'elle contient

5- Modéliser dans une ontologie le terrain et la nature des données associées

6- Créer un thésaurus aligné sur un référentiel pour la publication ou les données associées

7- Associer les deux approches dans une « **ontoterminologie** » (Almeida, 2017)

8- Prendre en compte l'évolutivité et l'incertitude… (Rabinowitz 2014)

HyperThésau Bibracte Numérique

Projets communs : publier et « ouvrir » des jeux-tests

1- Annuaire spatialisé des interventions sur le site de Bibracte

2- Emprise et résultats des prospections géophysiques sur le site de Bibracte

Dépôt sur Nakala → reprise par les portails cartographiques ArkeoGIS et ChronoCarto

3- Collection des objets archéologiques recueillis à Bibracte de 1867 à nos jours

Dépôt sur Nakala → reprise par la base collaborative en ligne Artefacts

4- Publications de/sur Bibracte de 1867 à nos jours et carnets de fouille du XIX^{e} s.

Transcription et documentarisation participatives des carnets de fouille manuscrits

Dépôt sur Persée → création d'une collection « Perséide »

HyperThésau Bibracte Numérique

« Vision » et objectifs communs

Comment « FAIR » : **F** = ID unique pérenne, métadonnées, diffusion (portails multiples)

Comment « FAIR » : **A** = à la machine et à l'Homme, protocoles, conditions d'accès, conservation à moyen et long terme

Comment « FAIR » : **I** = formats non-propriétaires, vocabulaire contrôlé, métadonnées

Comment « FAIR » : **R** = licences, description fine des données et de leur modèle conceptuel, données les plus « brutes » possibles (*raw data*) à chaque stade, « calculabilité » → réutilisation effective

Ce qui donne son sens de toute la démarche, c'est le « **R** ».

HyperThésau Bibracte Numérique

Infrastructure technique et épistémologie (organologie) :

« F » = dépôt sur un entrepôt public (p. ex. : Nakala) + métadonnées

« A » = moissonnage, reprise et accessibilité sur des portails (p. ex. : ArkeoGIS)

« I » = requêtes normalisées (structure des données et vocabulaire explicites) → identifier un *corpus* et y accéder

« R » = requêtes normalisées → extraire des *données* qualifiées et calculables

Toute structuration d'un jeu de données traduit une orientation, selon une problématique de recherche. **Paradoxe de l'interface** :

– elle ne *représente* pas les données, elle *présente un regard* sur elles

– elle réclame ou produit leur enrichissement selon le point de vue de la requête à venir

HyperThésau Bibracte Numérique

Percer la brume ou s'orienter

« Ouvrir » un jeu de données, c'est d'abord en fournir le **mode d'emploi**,
c.-à-d. sa description, au niveau le plus fin :

1. Son organisation (son **schéma conceptuel**),

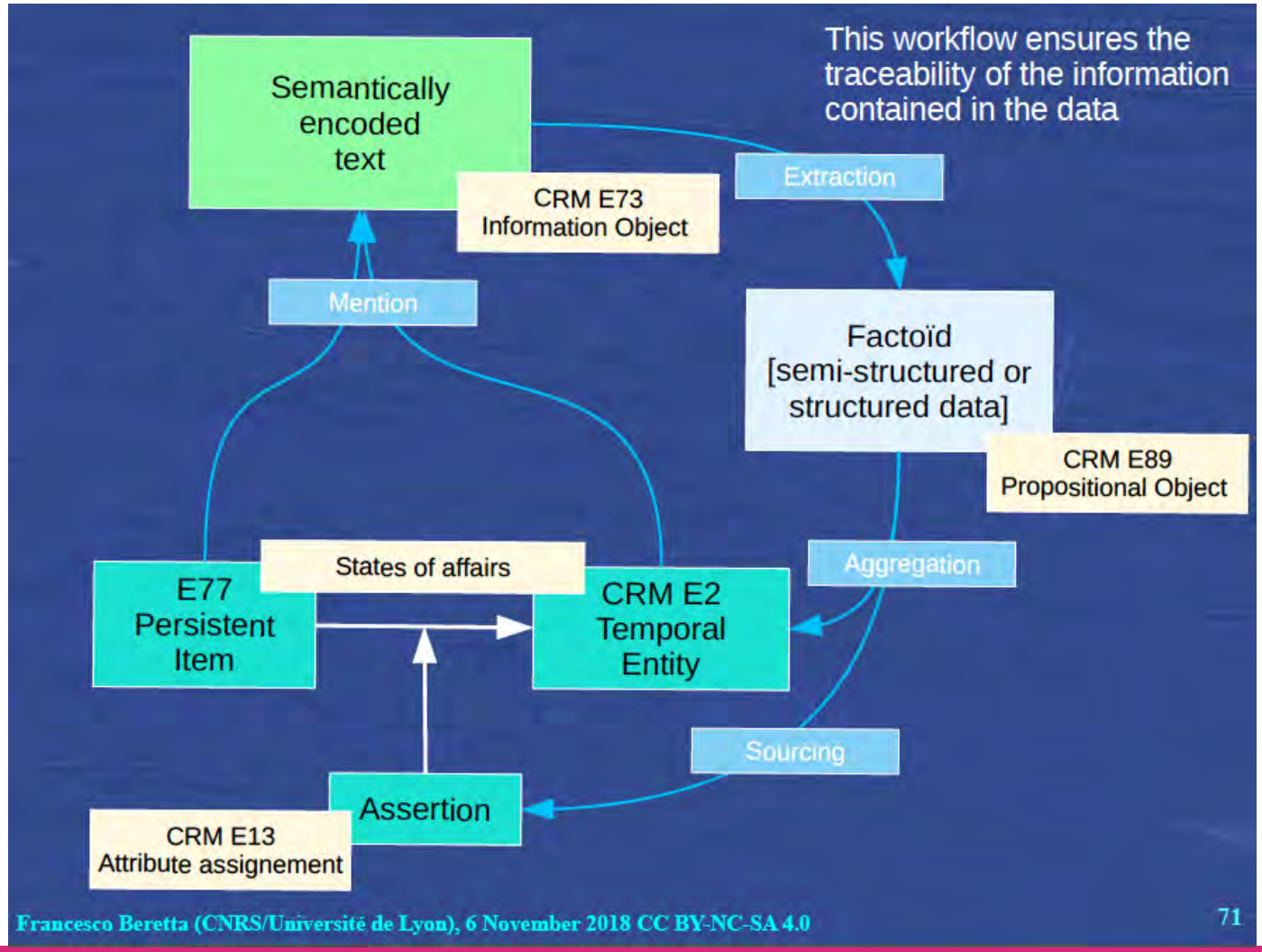
2. Son vocabulaire explicite et implicite (sa **terminologie**)

Voire :

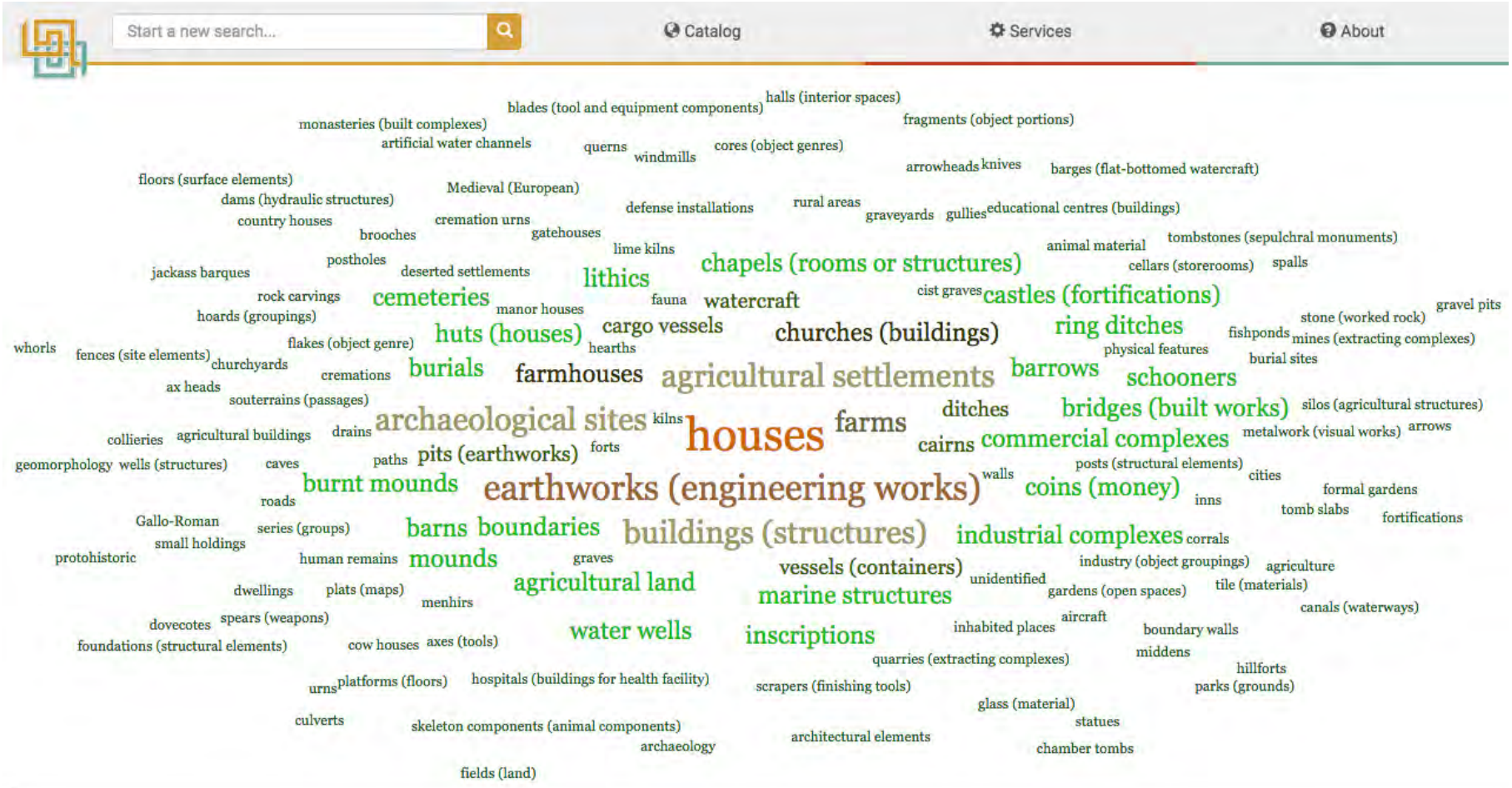
3. Les hypothèses discriminantes – p. ex. : périodisation – ayant présidé à sa construction (sa **contextualisation**)

L'ensemble est indispensable à sa compréhension en vue de sa **réutilisation**.

Data for History : extrait de modélisation (Beretta 2018)



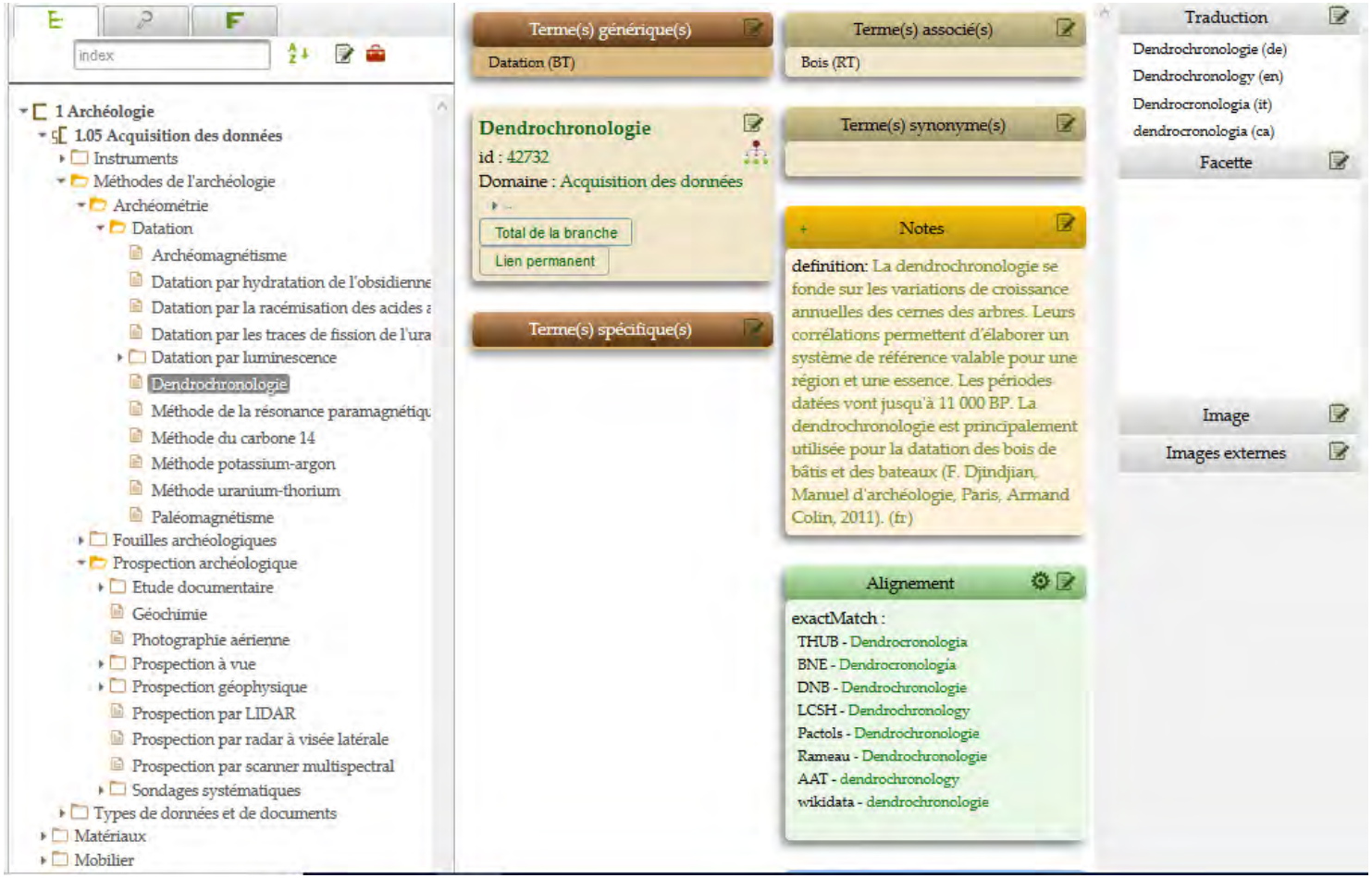
Vocabulaire-métier ou thésaurus ?



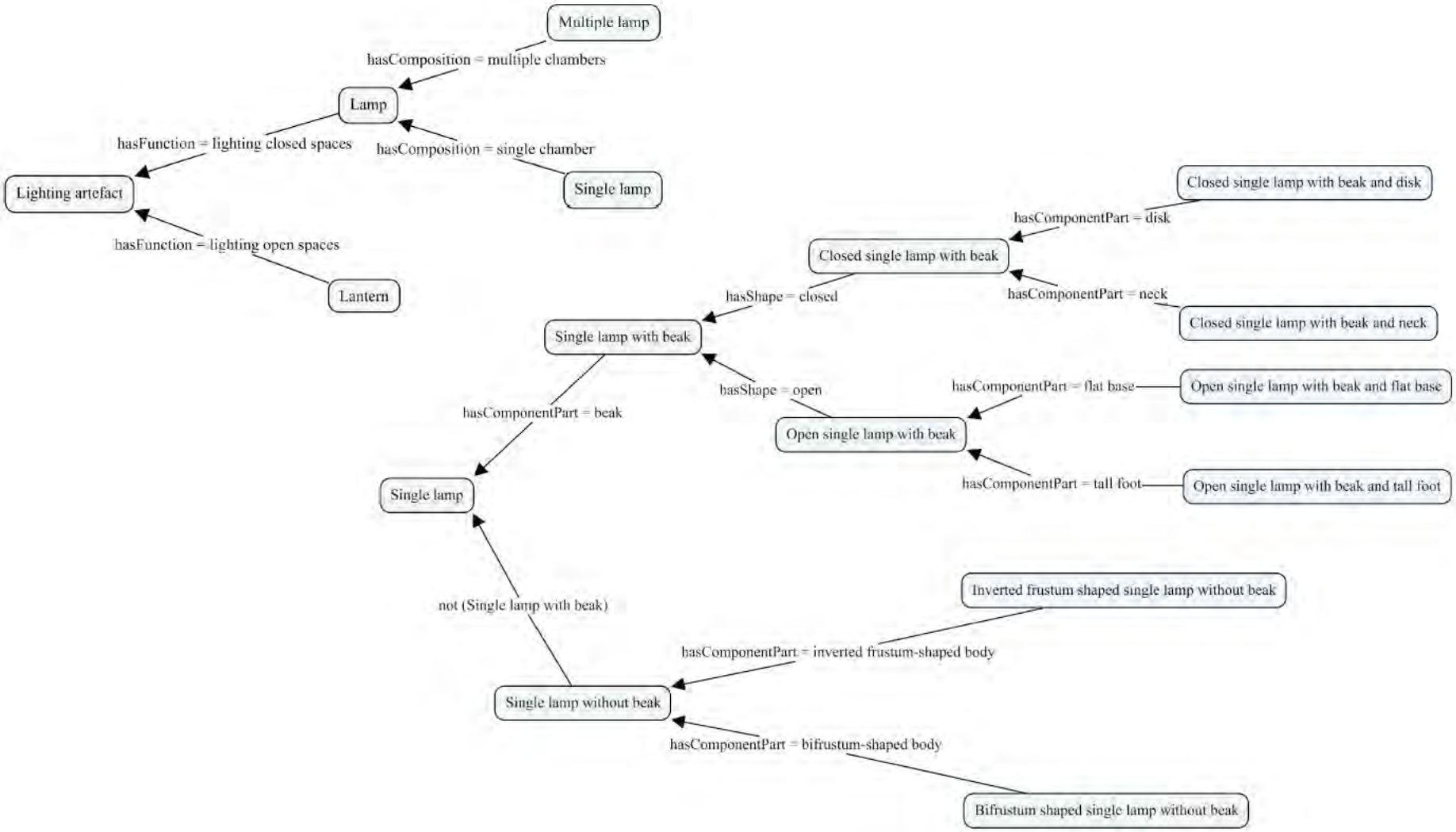
Nuage de tags ARIADNE, UE-FP7 programme 2012-2016

Vocabulaire *structuré* et *aligné*, *via* Opentheso (gestion de thésaurus)

« Dendrochronologie », construction d'un micro-thésaurus (travail en cours)



Ontologie et « ontoterminologie » (Almeida, Roche, Costa 2017)



HyperThésau Bibracte Numérique

Percer la brume (le phare)

Transformation des **enregistrements** du jeu de données : **description structurée en RDF**

– métadonnées « globales » (du jeu de données), en Dublin Core

– schéma conceptuel de la base de données (ontologie) en RDF

– vocabulaire contrôlé, aligné sur un thésaurus-pivot, lui-même aligné sur des référentiels du web sémantique, en SKOS

(Beretta 2018)

Les requêtes se font sur un *Sparql endpoint*.

Les résultats de la requête sont téléchargeables sous leur forme transformée et « calculable » (RDF).

HyperThésau Bibracte Numérique

S'orienter dans la brume (la boussole)

Paradigme du « **lac de données** » (pré-réparti en « étangs » de données cohérentes) :

– jeu de données dans son état/format originel

+ documenté par des métadonnées (Sawadogo 2019) :

– intrinsèques et « physiques » : propriétés techniques, quantitatives, d'identification (DC)

– intrinsèques et « logiques » : schéma conceptuel de la base de données, vocabulaire de référence utilisé, prévisualisation(s) calculée(s)

– globales : ontologie et thésaurus de référence, p. ex.

Un algorithme reconnaît, calcule et représente ces catégories de métadonnées pour permettre à l'utilisateur de pointer/qualifier les données identifiées par l'algorithme, puis relancer des itérations → **sélection/téléchargement des jeux de données originels**.

HyperThésau Bibracte Numérique

« Refuser de choisir, c'est encore choisir »

Quand on ne sait pas, on expérimente (et si on se trompe, on recommencera autrement)

1- **BibNum** : des objectifs pragmatiques, mais de vrais défis : grande masse de données, rétro-documentation sur un siècle, médiation sur site

2- **HyperThésau** : toutes les facettes de l'archéologie de terrain, en plusieurs langues, réunies en « lac », « étang » ou « marécage de données » : un « essai d'hydro-archéologie » ?!

3- **HisArc-RDF** : documentariser puis agréger des micro-corpus, « éparpillés façon puzzle »

4- « **Bibracte, Bulliot et moi** » : marier une frontière de recherche (fouille de textes manuscrit) et la « multitude », dans une expérience de documentarisation citoyenne

Bibliographie (très) sélective

Almeida B.,Roche C.,Costa R., « Archaeological classification and ontoterminology: the case of Islamic archaeology of the al-Andalus », Terminologie & Ontologie: Théories et Applications, TOTh 2017, pp.221-236 (https://hal.archives-ouvertes.fr/hal-01826942).

Beretta F., « Interoperability of historical data and FAIR principles : an ontology management environment (OntoME) for sharing and aligning data models », CLARIAH Seminar, 2018 (https://halshs.archives-ouvertes.fr/halshs-01975587).

Lukas D., Engel C., Mazzucato C., « Towards a Living Archive: Making Multi Layered Research Data and Knowledge Generation Transparent », *Journal of Field Archaeology*, vol. 43, 2018 (doi.org/10.1080/00934690.2018.1516110).

Rabinowitz A., « It's about time: historical periodization and Linked Ancient World Data » in *ISAW Papers* : Current Practice in Linked Open Data for the Ancient World, 7/22, 2014 (dlib.nyu.edu/awdl/isaw/isaw-papers/7/rabinowitz/).

Sawadogo P.N., Kibata T., Darmont J., « Metadata Management for Textual Documents in Data Lakes », 21st International Conference on Enterprise Information Systems (ICEIS 2019) (https://hal.archives-ouvertes.fr/hal-02012092).

HyperThésau Bibracte Numérique

Merci de votre attention (et de vos questions et suggestions)

Ce travail a été réalisé grâce au soutien financier du LABEX IMU (ANR-10-LABX-00) de l'Université de Lyon, dans le cadre du programme "Investissements d'Avenir" (ANR-11-IDEX-0007) géré par l'Agence Nationale de la Recherche (ANR).

BIBRACTE

Archéorient
Environnements et sociétés de l'Orient ancien

Présentation : Jean Pierre GIRARD (Archéorient)
Sébastien DUROST (Bibracte EPCC)