Les substrats numériques de la recherche : *documents, données, thésaurii, ontologies*

Cours de méthodologie

Nathalie Berton–Blivet & Thomas Bottini
nathalie.berton-blivet@cnrs.fr | thomas.bottini@cnrs.fr
Institut de Recherche en Musicologie
IReMus, UMR 8223 CNRS ✧ Sorbonne Université

Plan

[TB]

- Des idées aux données partagées
- Recenser & structurer les termes & les concepts avec un thésaurus

[NBB]

- Thésaurus & modélisation ontologique pour l'expression des connaissances historiques (+ *pause médiane quelque part…*)

[TB]

- Modéliser l'objet d'étude avec une ontologie

I — Des idées aux données partagées

Différence entre document et données

- Le vocable « données » dénote l'idée d'ensemble d'éléments de taille réduite structurés de manière similaire et prenant leur signification de la place qu'ils tiennent dans une série. Une donnée n'existe pas vraiment au singulier (*data*/*datum*).

Différence entre document et données

- Le vocable « données » dénote l'idée d'ensemble d'éléments de taille réduite structurés de manière similaire et prenant leur signification de la place qu'ils tiennent dans une série. Une donnée n'existe pas vraiment au singulier (*data*/*datum*).
- Le contenu d'un fichier Word ou PDF de type article ou mémoire ne relève pas de cette catégorie.

Différence entre document et données

- Le vocable « données » dénote l'idée d'ensemble d'éléments de taille réduite structurés de manière similaire et prenant leur signification de la place qu'ils tiennent dans une série. Une donnée n'existe pas vraiment au singulier (*data/datum*).
- Le contenu d'un fichier Word ou PDF de type article ou mémoire ne relève pas de cette catégorie.
- Un fichier Word ou PDF d'un catalogue pose question : il contient des données, mais celles-ci sont figées dans le texte, elles ne sont pas requêtables, appréhendables en sous-ensembles constitués par variation d'un paramètre de recherche. Autrement dit l'informatique ne sert à rien dans ce cas (ni dans sa capacité à formaliser, ni dans sa capacité à rechercher).

Différence entre document et données

- Le vocable « données » dénote l'idée d'ensemble d'éléments de taille réduite structurés de manière similaire et prenant leur signification de la place qu'ils tiennent dans une série. Une donnée n'existe pas vraiment au singulier (*data/datum*).
- Le contenu d'un fichier Word ou PDF de type article ou mémoire ne relève pas de cette catégorie.
- Un fichier Word ou PDF d'un catalogue pose question : il contient des données, mais celles-ci sont figées dans le texte, elles ne sont pas requêtables, appréhendables en sous-ensembles constitués par variation d'un paramètre de recherche. Autrement dit l'informatique ne sert à rien dans ce cas (ni dans sa capacité à formaliser, ni dans sa capacité à rechercher).
- fr.wikipedia.org/wiki/Abdel_Halim_Hafez *versus* www.wikidata.org/wiki/Q307786

Quand & pourquoi produire des données numériques ?

- Quand vous avez le sentiment que vous allez devenir fou/folle en faisant des tableaux dans Word.

Quand & pourquoi produire des données numériques ?

- Quand vous avez le sentiment que vous allez devenir fou/folle en faisant des tableaux dans Word.
- Quand vous accumulez des informations en série sur des objets de votre terrain qu'il vous faut analyser (requêter des sous-ensembles, faire des statistiques, construire des visualisations…).

Quand & pourquoi produire des données numériques ?

- Quand vous avez le sentiment que vous allez devenir fou/folle en faisant des tableaux dans Word.
- Quand vous accumulez des informations en série sur des objets de votre terrain qu'il vous faut analyser (requêter des sous-ensembles, faire des statistiques, construire des visualisations…).
- Quand vous sentez que l'effort de rédaction est distinct de l'effort de collecte, que les jeux de données constitués *ici et maintenant* pourraient répondre à de nouvelles questions de recherche *là bas et plus tard*.

Fréquentons un cas concret…

- *Catalogue des livres du roi (1660—1714/1725—1792)* (HDR de Laurence Decobert)
  - Première approche sous Word ?
  - Très vite, Excel !

Fréquentons un cas concret…

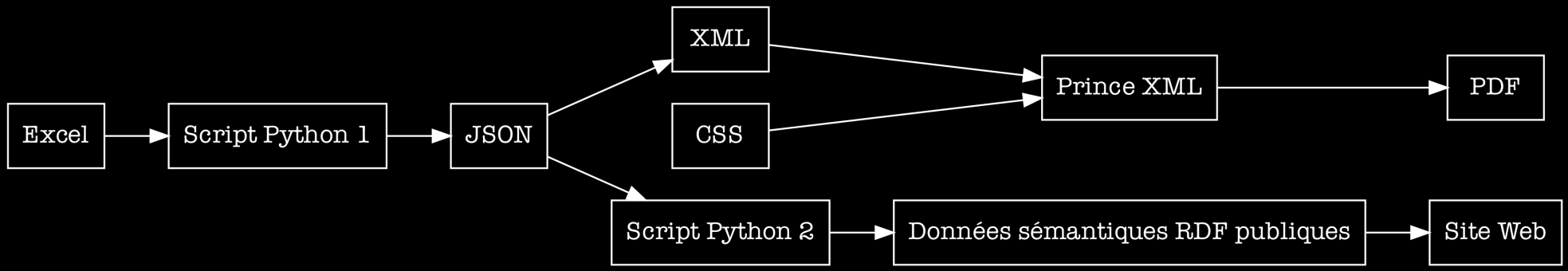
- *Catalogue des livres du roi (1660—1714/1725—1792)* (HDR de Laurence Decobert)
  - Première approche sous Word ?
  - Très vite, Excel !
  - Étudions la modélisation « implicite » dans les feuilles Excel. Les données sont structurées suivant un modèle, il faut identifier les types d'entités, leurs propriétés et les relations qu'elles entretiennent. Et rendre le tout explicite dans une structure informatique.

Fréquentons un cas concret…

- *Catalogue des livres du roi (1660—1714/1725—1792)* (HDR de Laurence Decobert)
  - Première approche sous Word ?
  - Très vite, Excel !
  - Étudions la modélisation « implicite » dans les feuilles Excel. Les données sont structurées suivant un modèle, il faut identifier les types d'entités, leurs propriétés et les relations qu'elles entretiennent. Et rendre le tout explicite dans une structure informatique.
  - Une première « interface » : le catalogue édité.
  - Projet de publication Web, à partir du même fichier de données.

Fréquentons un cas concret…

- *Catalogue des livres du roi (1660—1714/1725—1792)* (HDR de Laurence Decobert)
  - Première approche sous Word ?
  - Très vite, Excel !
  - Étudions la modélisation « implicite » dans les feuilles Excel. Les données sont structurées suivant un modèle, il faut identifier les types d'entités, leurs propriétés et les relations qu'elles entretiennent. Et rendre le tout explicite dans une structure informatique.
  - Une première « interface » : le catalogue édité.
  - Projet de publication Web, à partir du même fichier de données.



- cf. photo immonde

Problèmes épistémologiques

- Les données sont-elles vraiment données ? Les données ne sont-elles pas plutôt construites ?

Problèmes épistémologiques

- Les données sont-elles vraiment données ? Les données ne sont-elles pas plutôt construites ?
- Un jeu de données est-il réellement signifiant hors d'un contexte interprétatif donné ?

Problèmes épistémologiques

- Les données sont-elles vraiment données ? Les données ne sont-elles pas plutôt construites ?
- Un jeu de données est-il réellement signifiant hors d'un contexte interprétatif donné ?
- Le cas que nous venons d'étudier repose sur une myriade d'implicites : seul l'auteur de ce travail connaît le sens qu'il faut conférer aux colonnes et aux données de sa feuilles Excel.

ID	Attribution
0001	Josquin

➡ – Ne dit rien de l'activité de collecte de la donnée, du rapport à la source (aussi important que la donnée en elle-même si on parle de science).
– Ne dit rien du contexte argumentatif et du degré de certitude.
– Dans 10 ans, comment s'assurer que les chaînes de caractères correspondent bien à ce que l'auteur ou l'autrice avait en tête ?

Comment partager des données (Web)

- Contexte socio-technique : les données ouvertes et liées.
  - Échelle de qualité des données ouvertes de Tim Berners-Lee (Wikipedia)
  - https://5stardata.info/fr/ (rappel : projet Doremus).

Comment partager des données (Web)

- Contexte socio-technique : les données ouvertes et liées.
  - Échelle de qualité des données ouvertes de Tim Berners-Lee (Wikipedia)
  - https://5stardata.info/fr/ (rappel : projet Doremus).
- Les trois grandes opérations informationnelles à réaliser pour passer d'un fichier Excel personnel à un jeu de données « compréhensible » sur le Web :
  - Transformer les identifiants en URI.
  - Aligner les entités identifiées (personnes, lieux, notions…) sur des référentiels et des thésaurii, principalement pour des raisons de désambiguïsation.
  - Feuilles & colonnes (modèle) => choisir une ontologie adéquate au corpus, autour de laquelle existe une communauté de pratiques qui réfléchit et formalise des invariants dans la manière de considérer leurs objets d'étude.

II – Thésaurii

Constituer un thésaurus

- Présentation d'un projet concret en cours : indexation du corpus d'estampes du Mercure galant 1678—1710 (IReMus/ObTIC).

Constituer un thésaurus

- Présentation d'un projet concret en cours : indexation du corpus d'estampes du Mercure galant 1678—1710 (IReMus/ObTIC).
- Organisation, description et indexation du corpus.

Constituer un thésaurus

- Présentation d'un projet concret en cours : indexation du corpus d'estampes du Mercure galant 1678—1710 (IReMus/ObTIC).
- Organisation, description et indexation du corpus.
- Organisation du thésaurus.

Constituer un thésaurus

- Présentation d'un projet concret en cours : indexation du corpus d'estampes du Mercure galant 1678—1710 (IReMus/ObTIC).
- Organisation, description et indexation du corpus.
- Organisation du thésaurus.
- Alignement sur des thésaurii mondiaux
  - Patrimoine, en général : Getty AAT
  - Histoire de l'art : Iconclass

Constituer un thésaurus

- Présentation d'un projet concret en cours : indexation du corpus d'estampes du Mercure galant 1678—1710 (IReMus/ObTIC).
- Organisation, description et indexation du corpus.
- Organisation du thésaurus.
- Alignement sur des thésaurii mondiaux
  - Patrimoine, en général : Getty AAT
  - Histoire de l'art : Iconclass
- Synergie scientifique (une idée : le projet Emblematica)

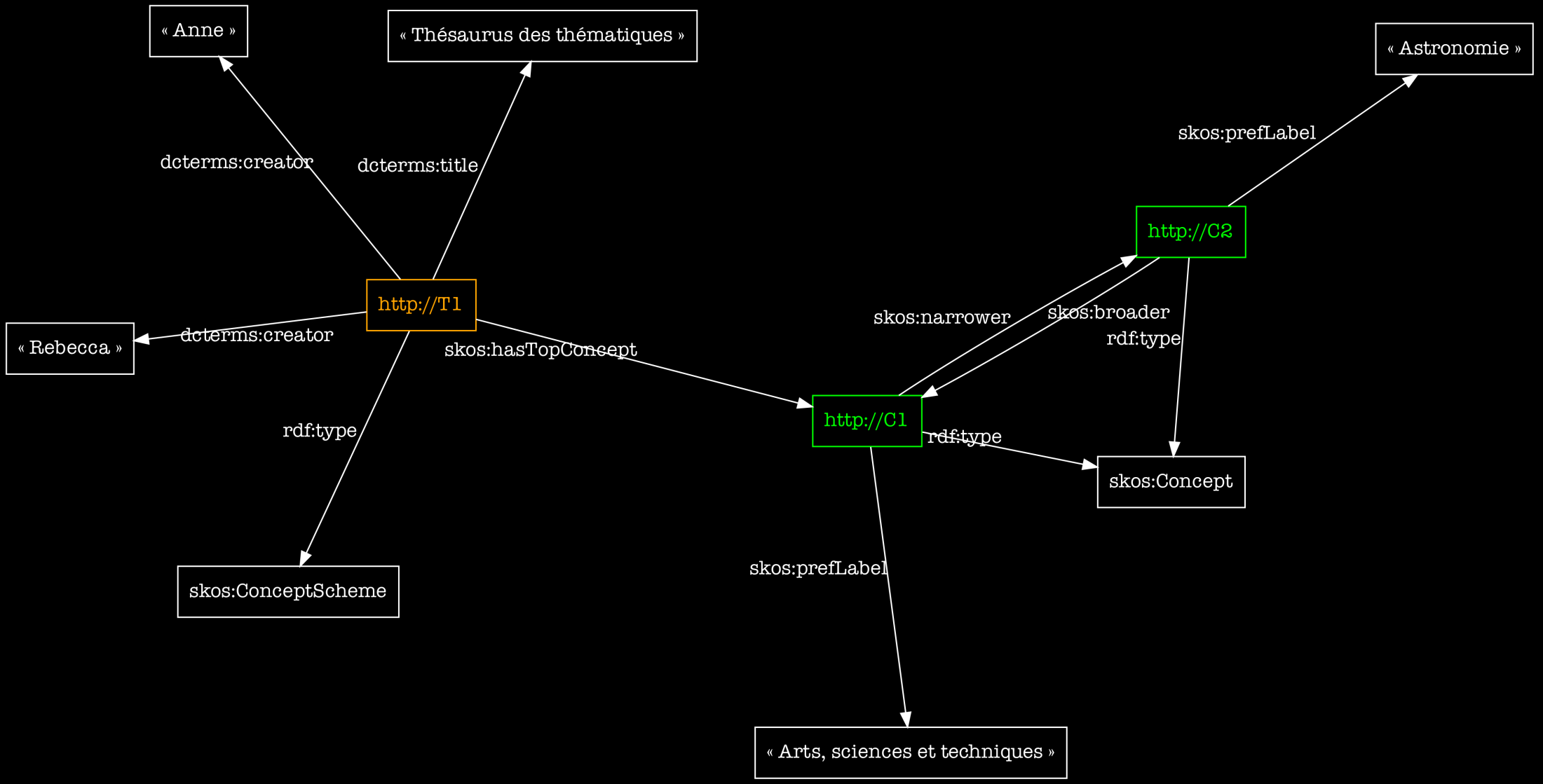
Constituer un thésaurus

- Présentation d'un projet concret en cours : indexation du corpus d'estampes du Mercure galant 1678—1710 (IReMus/ObTIC).
- Organisation, description et indexation du corpus.
- Organisation du thésaurus.
- Alignement sur des thésaurii mondiaux
  - Patrimoine, en général : Getty AAT
  - Histoire de l'art : Iconclass
- Synergie scientifique (une idée : le projet Emblematica)
- Écueil épistémologique potentiel : confrontation entre approche *bottom-up* (localisée + située) et *top-down* (globalisée + détachée des pratiques effectives).

Constituer un thésaurus

- Présentation d'un projet concret en cours : indexation du corpus d'estampes du Mercure galant 1678—1710 (IReMus/ObTIC).
- Organisation, description et indexation du corpus.
- Organisation du thésaurus.
- Alignement sur des thésaurii mondiaux
  - Patrimoine, en général : Getty AAT
  - Histoire de l'art : Iconclass
- Synergie scientifique (une idée : le projet Emblematica)
- Écueil épistémologique potentiel : confrontation entre approche *bottom-up* (localisée + située) et *top-down* (globalisée + détachée des pratiques effectives).
- Formaliser un thésaurus avec SKOS (*Simple Knowledge Organization System*).

1 thésaurus & 2 concepts en SKOS



Thésaurus & modélisation ontologique pour l'expression des connaissances historiques

Modéliser l'objet d'étude avec une ontologie

Sémantique des connaissances historiques

- **Personnes**
  - Représenter les propriétés d'un individu historique (patronyme, pseudonyme, dates, relations de parenté…).

Sémantique des connaissances historiques

- **Personnes**
  - Représenter les propriétés d'un individu historique (patronyme, pseudonyme, dates, relations de parenté…).
  - Représenter ses activités & relations sociales.

Sémantique des connaissances historiques

- **Personnes**
  - Représenter les propriétés d'un individu historique (patronyme, pseudonyme, dates, relations de parenté…).
  - Représenter ses activités & relations sociales.
- **Lieux**
  - Inclusion administrative ≠ inclusion géographique ≠ inclusion temporelle

Sémantique des connaissances historiques

- **Personnes**
  - Représenter les propriétés d'un individu historique (patronyme, pseudonyme, dates, relations de parenté…).
  - Représenter ses activités & relations sociales.
- **Lieux**
  - Inclusion administrative ≠ inclusion géographique ≠ inclusion temporelle
  - Représenter la diachronie (ex : Ris + Orangis = Ris-Orangis (1793)).

Sémantique des connaissances historiques

- **Personnes**
  - Représenter les propriétés d'un individu historique (patronyme, pseudonyme, dates, relations de parenté…).
  - Représenter ses activités & relations sociales.
- **Lieux**
  - Inclusion administrative ≠ inclusion géographique ≠ inclusion temporelle
  - Représenter la diachronie (ex : Ris + Orangis = Ris-Orangis (1793)).
- **Institutions**
  - Les sémantiques de la relation unissant « Académie » et « Académie de Besançon » et de la relation unissant « Académie de Besançon » et « Associé-résident de l'Académie de Besançon » diffèrent.

Sémantique des connaissances historiques

- **Personnes**
  - Représenter les propriétés d'un individu historique (patronyme, pseudonyme, dates, relations de parenté…).
  - Représenter ses activités & relations sociales.
- **Lieux**
  - Inclusion administrative ≠ inclusion géographique ≠ inclusion temporelle
  - Représenter la diachronie (ex : Ris + Orangis = Ris-Orangis (1793)).
- **Institutions**
  - Les sémantiques de la relation unissant « Académie » et « Académie de Besançon » et de la relation unissant « Académie de Besançon » et « Associé-résident de l'Académie de Besançon » diffèrent.
  - Représenter le fait que « Pierre Beauchamp était membre de l'Académie royale de Danse en 1680 » ?

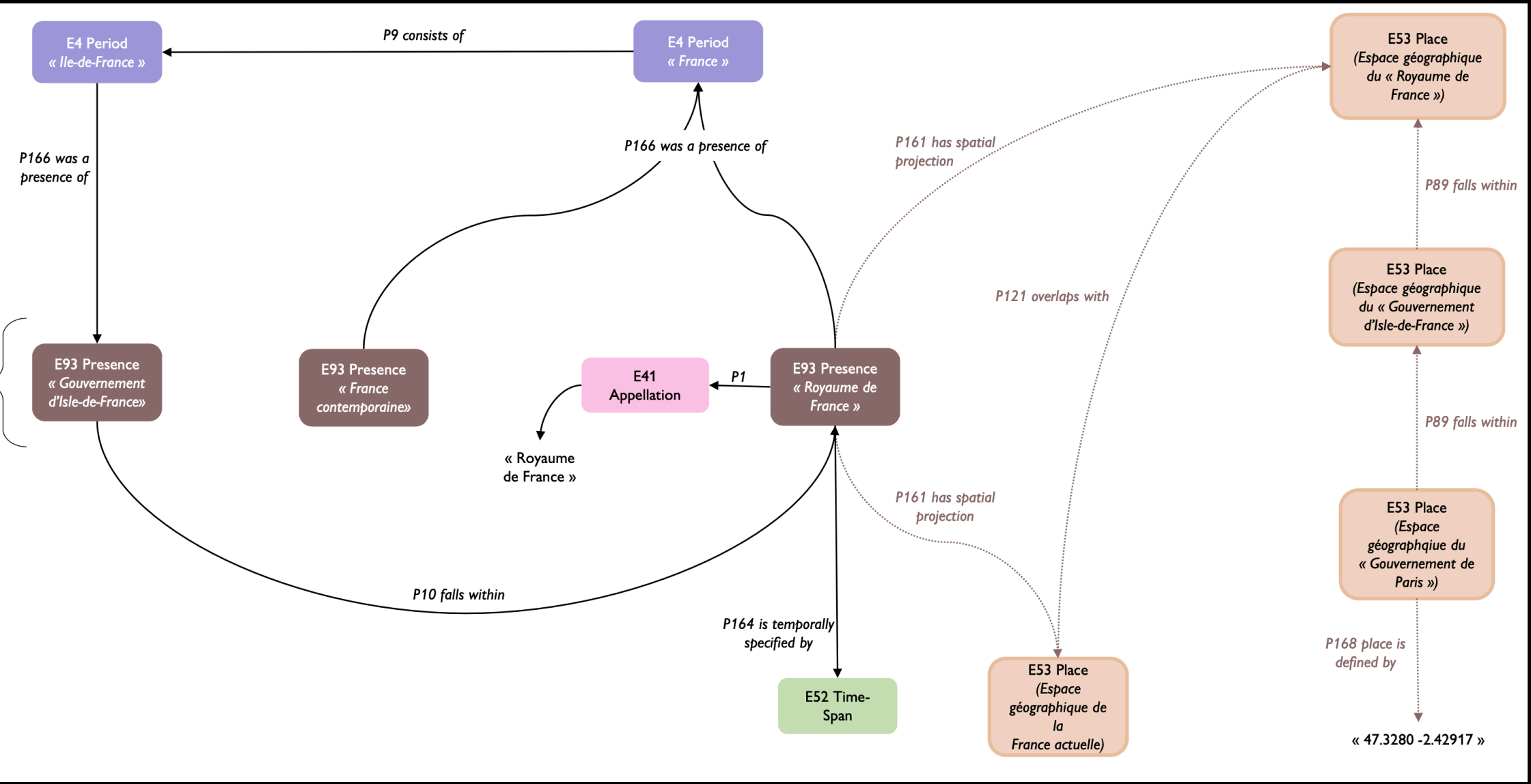
Une question, une ontologie

- L'ontologie SKOS concerne la définition des concepts (`skos:definition`, `skos:prefLabel`...) et des rapports qu'ils entretiennent (`skos:narrower`, `skos:broader`, `skos:related`...). Les entités historiques ne sont pas vraiment des concepts.

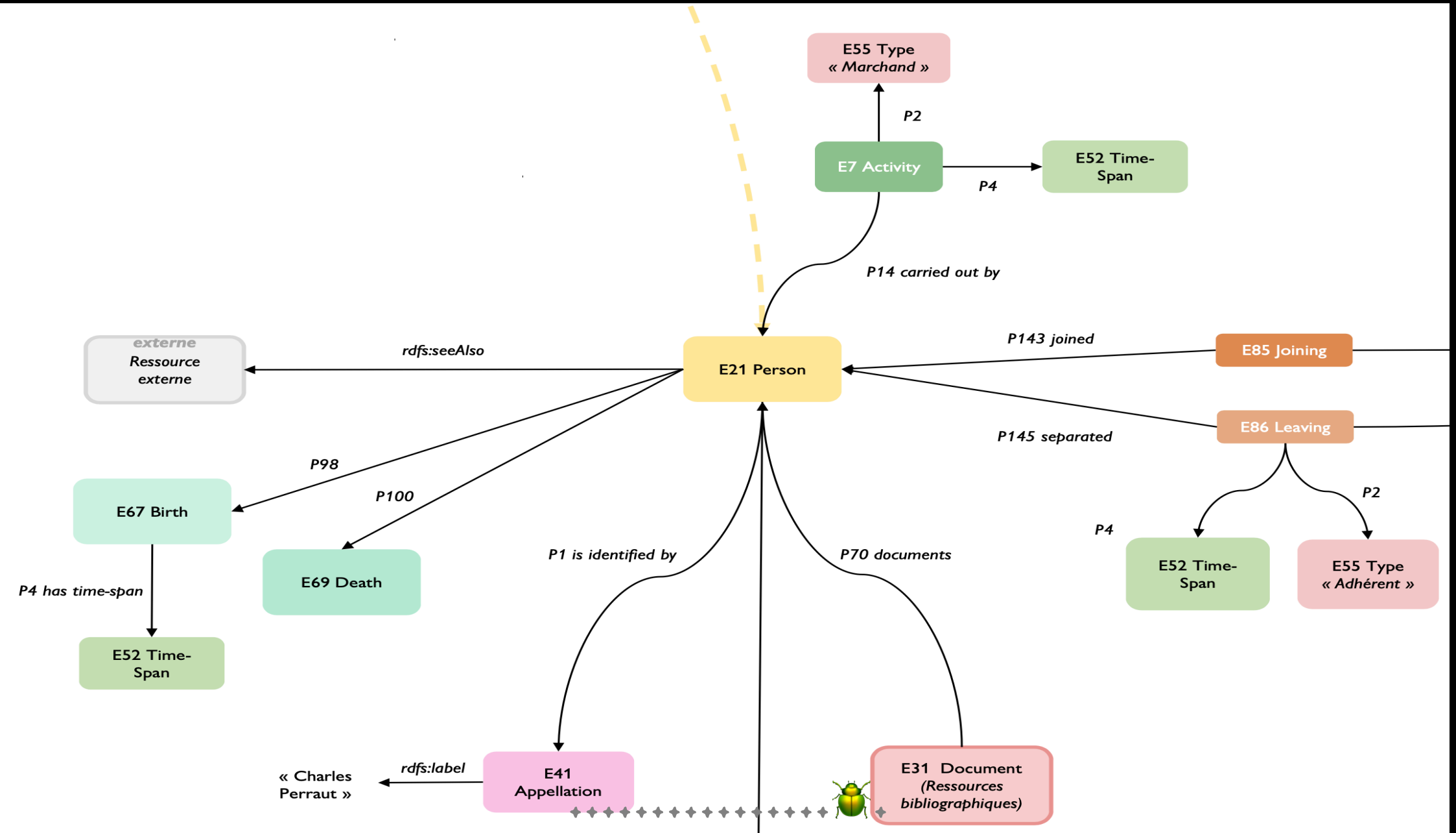
Une question, une ontologie

- L'ontologie SKOS concerne la définition des concepts (`skos:definition`, `skos:prefLabel`...) et des rapports qu'ils entretiennent (`skos:narrower`, `skos:broader`, `skos:related`...). Les entités historiques ne sont pas vraiment des concepts.
- Ontologie adaptée à l'expression de connaissances historiques (par exemple, le CIDOC-CRM). Quelques exemples...

Lieux historiques avec le CRM



Personnes historiques avec le CRM



Groupes historiques avec le CRM