Formations autour de la diffusion de la recherche et de la science ouverte
Les humanités numériques en pratique
15 février 2024

MODÉLISER LES CONNAISSANCES EN SHS : POURQUOI, COMMENT, JUSQU'OÙ ?

… avec l'ontologie cidoc-crm !

Thomas Bottini
thomas.bottini@cnrs.fr
Institut de Recherche en Musicologie — IReMus, UMR 8223 CNRS

PLAN

La figure du modélisateur/de la modélisatrice dans un projet de SHS « FAIR »

Quelques mots sur le Web sémantique comme milieu technique adapté à l'expression et à la diffusion des données de la recherche

Les fondements de l'ontologie CIDOC-CRM

Cas concrets récurrents

DE L'ACTIVITÉ DE RECHERCHE À LA DONNÉE

Comment « l'ingénierie des données » intervient-elle aux différentes étapes d'un projet de recherche en SHS ?

Comment bien faire du FAIR ?

ANALYSER

- Les chercheurs et chercheuses en situation de production de connaissances ont des degrés de réflexivité différents sur leurs pratiques de génération de données informatiques structurées. Le premier rôle de l'ingénieur est alors d'ordre maïeutique :
  - Il faut poser des questions (entretiens d'explicitation), confronter le chercheur ou la chercheuse à des cas limites pour l'amener à mieux comprendre ses objets d'étude, et parfois même ses manières de les questionner.
  - L'ingénieur aide donc à révéler la structure interne des sources et des phénomènes étudiés : une dimension heuristique s'ajoute.

ANALYSER

- Les chercheurs et chercheuses en situation de production de connaissances ont des degrés de réflexivité différents sur leurs pratiques de génération de données informatiques structurées. Le premier rôle de l'ingénieur est alors d'ordre maïeutique :
  - Il faut poser des questions (entretiens d'explicitation), confronter le chercheur ou la chercheuse à des cas limites pour l'amener à mieux comprendre ses objets d'étude, et parfois même ses manières de les questionner.
  - L'ingénieur aide donc à révéler la structure interne des sources et des phénomènes étudiés : une dimension heuristique s'ajoute.
- En SHS, ces connaissances peuvent résulter d'une activité descriptive (sources) ou interprétative (choses dites à propos des sources).
- Il faut analyser à la fois les objets et les produits de l'activité analytique (quoi ?), ainsi que la situation de cette activité, son contexte socio-technico-épistémologique (qui ? comment ? pourquoi ?).

MODÉLISER

- Quoi ?
  - Les connaissances nouvelles s'incarnant dans des « données scientifiques ».
  - Les sources auxquelles elles se rapportent.
  - Les faits sociaux, les concepts, les objets matériels, les objets sémiotiques… dénotés ou connotés par les sources et/ou observables sur le terrain.
  - Le contexte de production des connaissances (qu'est ce qui fait qu'une donnée est « scientifique » ?).

MODÉLISER

- Quoi ?
  - Les connaissances nouvelles s'incarnant dans des « données scientifiques ».
  - Les sources auxquelles elles se rapportent.
  - Les faits sociaux, les concepts, les objets matériels, les objets sémiotiques… dénotés ou connotés par les sources et/ou observables sur le terrain.
  - Le contexte de production des connaissances (qu'est ce qui fait qu'une donnée est « scientifique » ?).
- Dès lors, avoir un modèle conceptuel de type ontologique en tête en amont de la phase d'analyse permet d'organiser l'écoute et l'analyse des pratiques de production de connaissances.

MODÉLISER

- Quoi ?
  - Les connaissances nouvelles s'incarnant dans des « données scientifiques ».
  - Les sources auxquelles elles se rapportent.
  - Les faits sociaux, les concepts, les objets matériels, les objets sémiotiques… dénotés ou connotés par les sources et/ou observables sur le terrain.
  - Le contexte de production des connaissances (qu'est ce qui fait qu'une donnée est « scientifique » ?).
- Dès lors, avoir un modèle conceptuel de type ontologique en tête en amont de la phase d'analyse permet d'organiser l'écoute et l'analyse des pratiques de production de connaissances.
- Modéliser est une situation de travail typique des HN dans laquelle l'ingénierie n'a pas un rôle ancillaire (cf. *supra* fonction heuristique/ épistémologique). En définissant formellement les objets convoqués par la recherche, l'ingénieur•e participe à leur constitution.

GÉNÉRER, GÉRER & DIFFUSER

Les challenges techniques sont de taille :

- Pour la saisie, l'idéal est de mettre en place des interfaces de saisie ergonomiques…
- …mais c'est là où « le bât blesse » : les ontologies sont souples, riches, expressives (les données qu'elles modélisent se donnent sous la forme de graphes ouverts), mais pas instrumentées pour l'édition directe par des humains par rapport aux données relationnelles SQL, de nature plus « tabulaire » et qui s'éditent naturellement avec des formulaires.

GÉNÉRER, GÉRER & DIFFUSER

Les challenges techniques sont de taille :

- Pour la saisie, l'idéal est de mettre en place des interfaces de saisie ergonomiques…
- …mais c'est là où « le bât blesse » : les ontologies sont souples, riches, expressives (les données qu'elles modélisent se donnent sous la forme de graphes ouverts), mais pas instrumentées pour l'édition directe par des humains par rapport aux données relationnelles SQL, de nature plus « tabulaire » et qui s'éditent naturellement avec des formulaires.
- Reprendre les données existantes, les rendre conformes à l'ontologie retenue.

GÉNÉRER, GÉRER & DIFFUSER

Les challenges techniques sont de taille :

- Pour la saisie, l'idéal est de mettre en place des interfaces de saisie ergonomiques…
- …mais c'est là où « le bât blesse » : les ontologies sont souples, riches, expressives (les données qu'elles modélisent se donnent sous la forme de graphes ouverts), mais pas instrumentées pour l'édition directe par des humains par rapport aux données relationnelles SQL, de nature plus « tabulaire » et qui s'éditent naturellement avec des formulaires.
- Reprendre les données existantes, les rendre conformes à l'ontologie retenue.
- Créer un accès pour les machines (API, SPARQL endpoint).

GÉNÉRER, GÉRER & DIFFUSER

Les challenges techniques sont de taille :

- Pour la saisie, l'idéal est de mettre en place des interfaces de saisie ergonomiques…
- …mais c'est là où « le bât blesse » : les ontologies sont souples, riches, expressives (les données qu'elles modélisent se donnent sous la forme de graphes ouverts), mais pas instrumentées pour l'édition directe par des humains par rapport aux données relationnelles SQL, de nature plus « tabulaire » et qui s'éditent naturellement avec des formulaires.
- Reprendre les données existantes, les rendre conformes à l'ontologie retenue.
- Créer un accès pour les machines (API, SPARQL endpoint).
- Définir une politique pour la publication des données représentant des concepts ou des termes dans des vocabulaires contrôlés/thésauri et des données représentant des entités.

GÉNÉRER, GÉRER & DIFFUSER

Les challenges techniques sont de taille :

- Pour la saisie, l'idéal est de mettre en place des interfaces de saisie ergonomiques…
- …mais c'est là où « le bât blesse » : les ontologies sont souples, riches, expressives (les données qu'elles modélisent se donnent sous la forme de graphes ouverts), mais pas instrumentées pour l'édition directe par des humains par rapport aux données relationnelles SQL, de nature plus « tabulaire » et qui s'éditent naturellement avec des formulaires.
- Reprendre les données existantes, les rendre conformes à l'ontologie retenue.
- Créer un accès pour les machines (API, SPARQL endpoint).
- Définir une politique pour la publication des données représentant des concepts ou des termes dans des vocabulaires contrôlés/thésauri et des données représentant des entités.
- Créer des interfaces de consultation pour les humains (sites Web).

VOCABULAIRE DE BASE

LE WEB SÉMANTIQUE, EN UNE SLIDE

- Promesse d'une base de données à l'échelle du Web. Le Web initial (Tim Berners Lee, 1991) était un Web de documents liés (hypertexte), le Web sémantique est une Web de données liées.

LE WEB SÉMANTIQUE, EN UNE SLIDE

- Promesse d'une base de données à l'échelle du Web. Le Web initial (Tim Berners Lee, 1991) était un Web de documents liés (hypertexte), le Web sémantique est une Web de données liées.
- Chaque donnée est identifiée par une URL.

LE WEB SÉMANTIQUE, EN UNE SLIDE

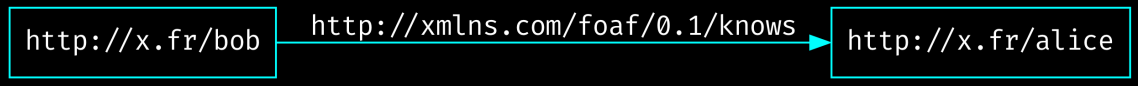
- Promesse d'une base de données à l'échelle du Web. Le Web initial (Tim Berners Lee, 1991) était un Web de documents liés (hypertexte), le Web sémantique est une Web de données liées.
- Chaque donnée est identifiée par une URL.
- Toute information s'exprime sous la forme d'un triplet (sujet/prédicat/objet) dans un langage de description, le RDF.

LE WEB SÉMANTIQUE, EN UNE SLIDE

- Promesse d'une base de données à l'échelle du Web. Le Web initial (Tim Berners Lee, 1991) était un Web de documents liés (hypertexte), le Web sémantique est une Web de données liées.
- Chaque donnée est identifiée par une URL.
- Toute information s'exprime sous la forme d'un triplet (sujet/prédicat/objet) dans un langage de description, le RDF.
- La connexion de ces triplets RDF forme un graphe.

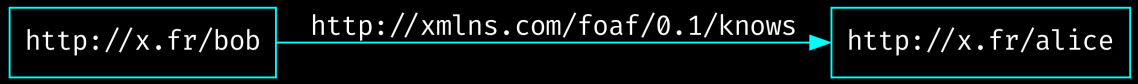
LE WEB SÉMANTIQUE, EN UNE SLIDE

- Promesse d'une base de données à l'échelle du Web. Le Web initial (Tim Berners Lee, 1991) était un Web de documents liés (hypertexte), le Web sémantique est une Web de données liées.
- Chaque donnée est identifiée par une URL.
- Toute information s'exprime sous la forme d'un triplet (sujet/prédicat/objet) dans un langage de description, le RDF.
- La connexion de ces triplets RDF forme un graphe.
- Chaque prédicat est également identifié par une URL.



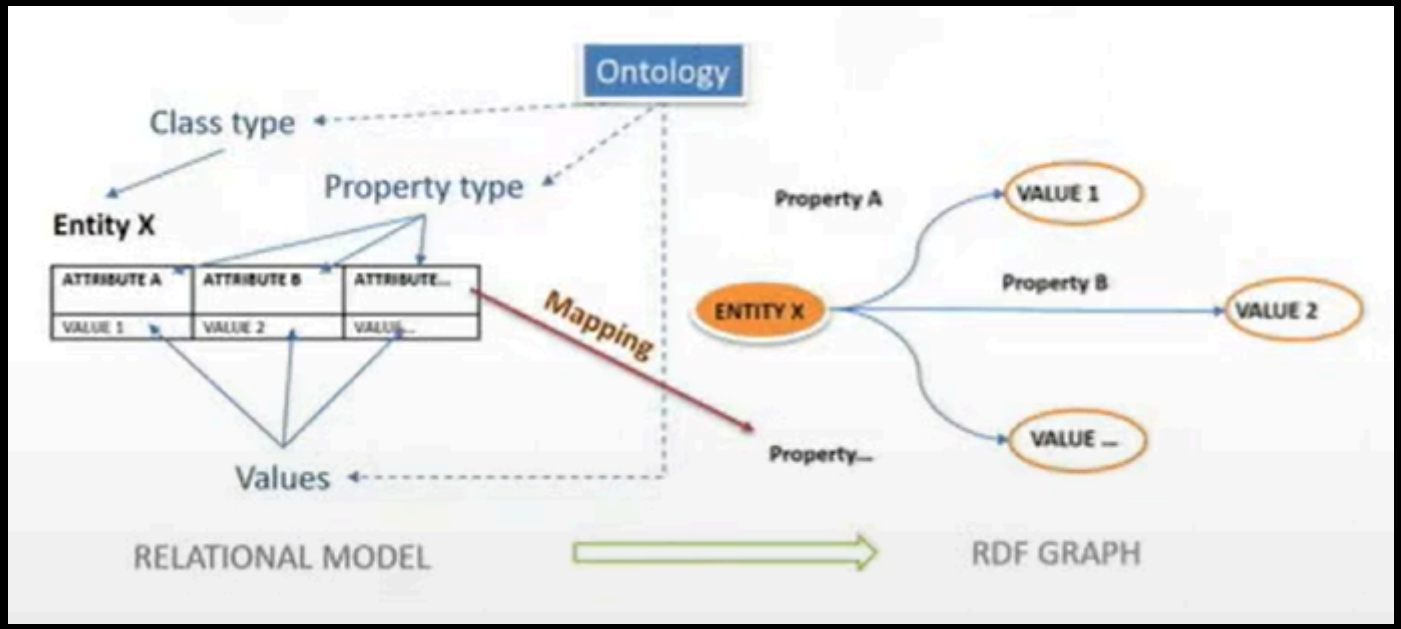
LE WEB SÉMANTIQUE, EN UNE SLIDE

- Promesse d'une base de données à l'échelle du Web. Le Web initial (Tim Berners Lee, 1991) était un Web de documents liés (hypertexte), le Web sémantique est une Web de données liées.
- Chaque donnée est identifiée par une URL.
- Toute information s'exprime sous la forme d'un triplet (sujet/prédicat/objet) dans un langage de description, le RDF.
- La connexion de ces triplets RDF forme un graphe.
- Chaque prédicat est également identifié par une URL.



- C'est le milieu technique idéal pour des données FAIR.

DONNÉES RELATIONNELLES VS GRAPHE RDF



Corago in LOD - Seminar by Angelo Pompilio and Paolo Bonora, Digital Humanities and Digital Knowledge, Università di Bologna, 2017.

QU'EST CE QU'UNE ONTOLOGIE ?

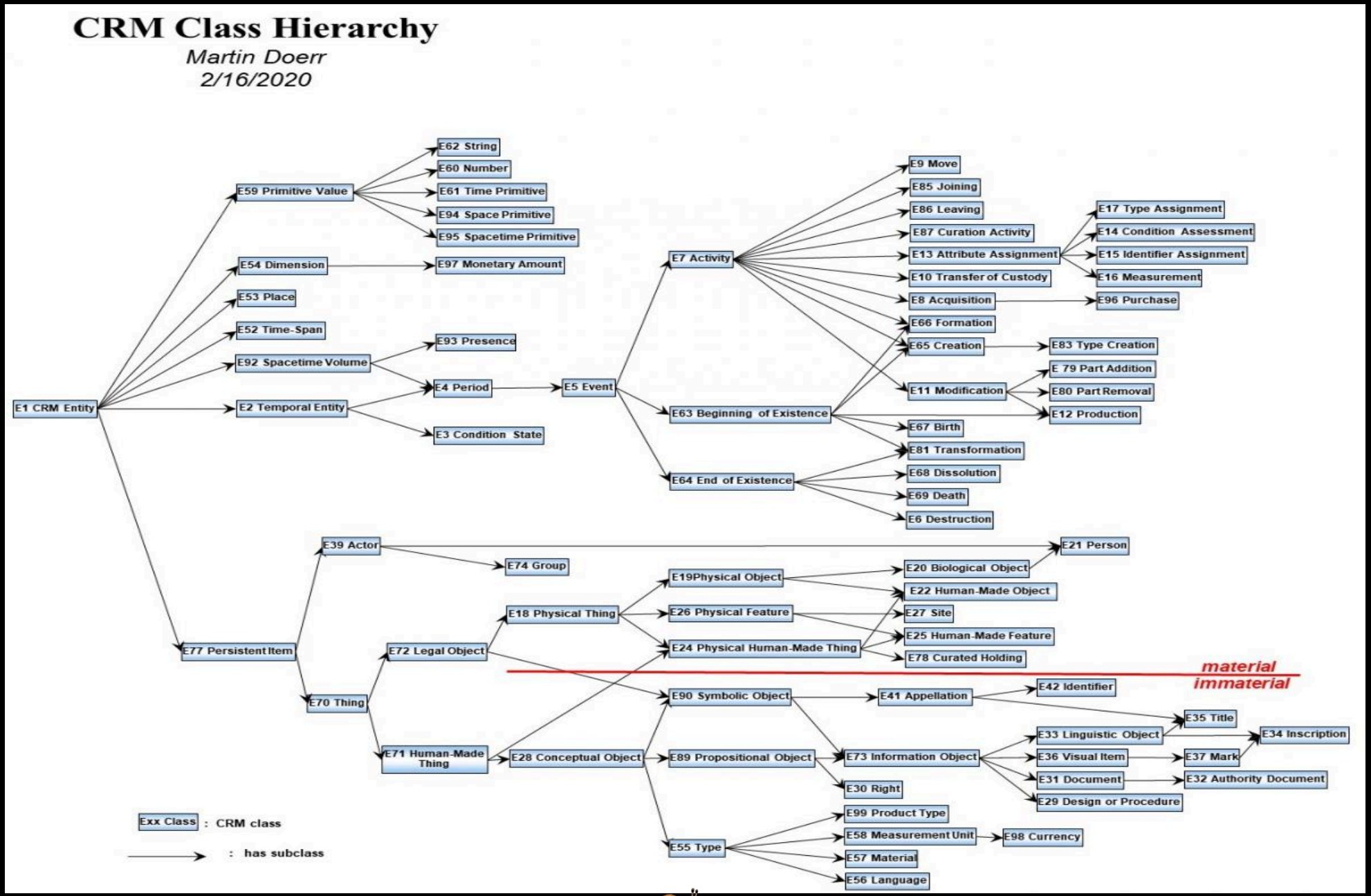
- Formalisation d'un modèle conceptuel pour un domaine identifié proposant des :
  - Classes : types d'entités peuplant le domaine, possiblement organisées selon des relations d'héritage (spécificité). On appelle individu une ressource qui est du type d'une classe.
  - Propriétés : aspects, caractéristiques, attributs possibles de ces classes, qui peuvent soit pointer vers une valeur, soit vers un individu.
- Utiliser les classes et les propriétés d'une ontologie confère ainsi une sémantique partagée aux données RDF (les individus identifiés par des URL seront des sujets ou des objets, les propriétés des classes seront des prédicats).
- Vous connaissez peut-être déjà une ontologie : SKOS (pour construire des thésauri).

LE CIDOC-CRM

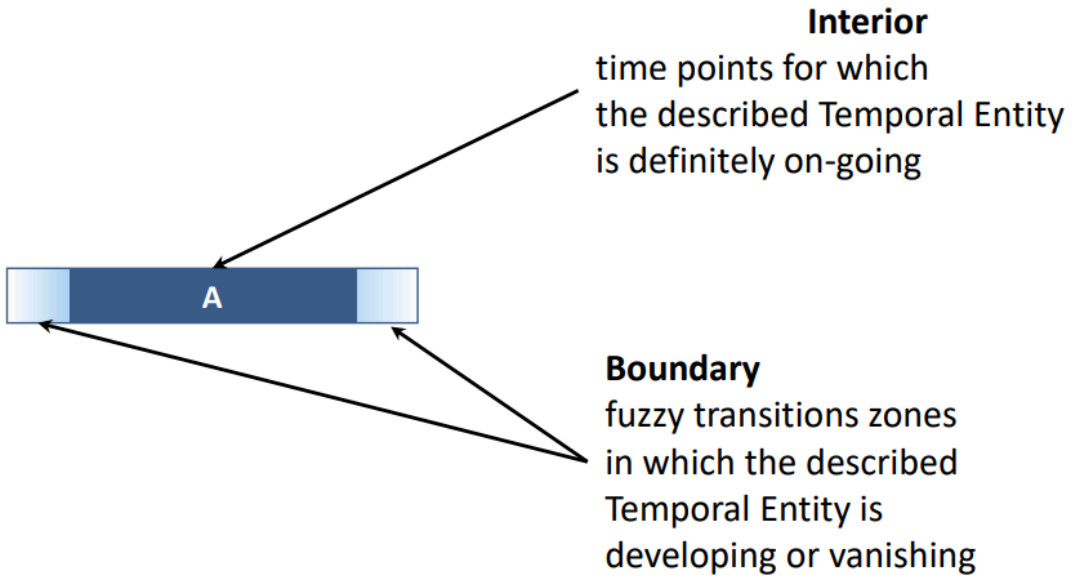
LE CIDOC-CRM EN BREF

- Le CIDOC-CRM est une ontologie qui documente le patrimoine matériel et immatériel ainsi que les processus de production de connaissances à son propos.
- https://www.cidoc-crm.org/
- Venant du monde des musées, elle est désormais utilisée dans tous les domaines des HN.
- Elle est extrêmement abstraite et générique.
- Ontologie centrée événement (nous y reviendrons dans les exemples...)
- Classes et propriétés : https://cidoc-crm.org/html/cidoc_crm_v7.1.2.html

HIÉRARCHIE DES CLASSES



LE TEMPS DANS LE CRM



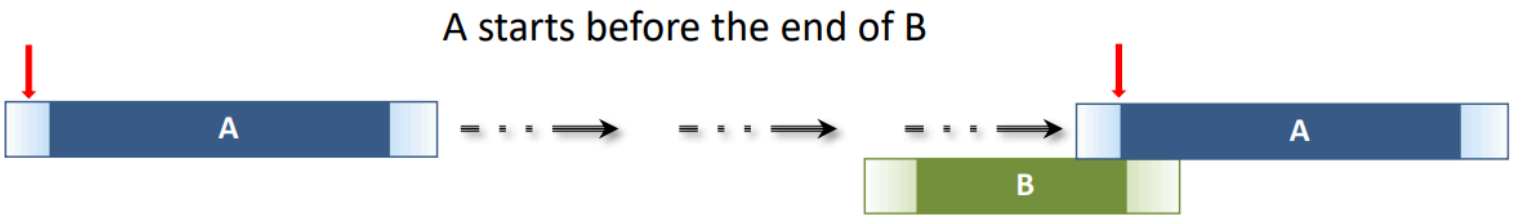


Figure 7: Explanation of Interior and Boundary and an Example of Use from P174 starts before the end of (ends after the start of).

CAS CONCRETS RÉCURRENTS

CONVENTIONS

Dans les exemples, on utilisera les préfixes suivants pour les URL :

```
PREFIX crm: <http://www.cidoc-crm.org/cidoc-crm/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX su:  <http://www.sorbonne-universite.fr/id/>
```

Ainsi, l'URL :

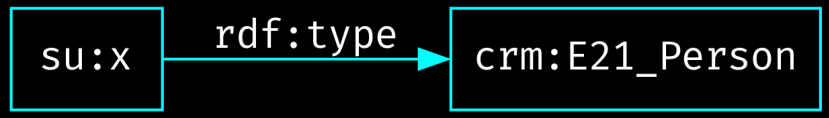
```
http://www.sorbonne-universite.fr/id/tralala
```

sera notée :

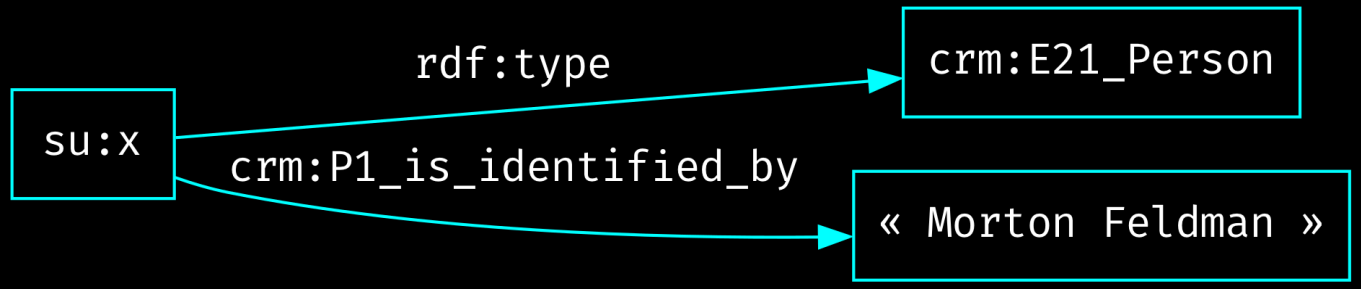
```
su:tralala
```

...ce qui évitera d'avoir des exemples illisibles, mais il ne faudra pas perdre de vue que toute ressource (en incluant les classes et les propriétés du CRM) est identifiée par une URI sur le Web.

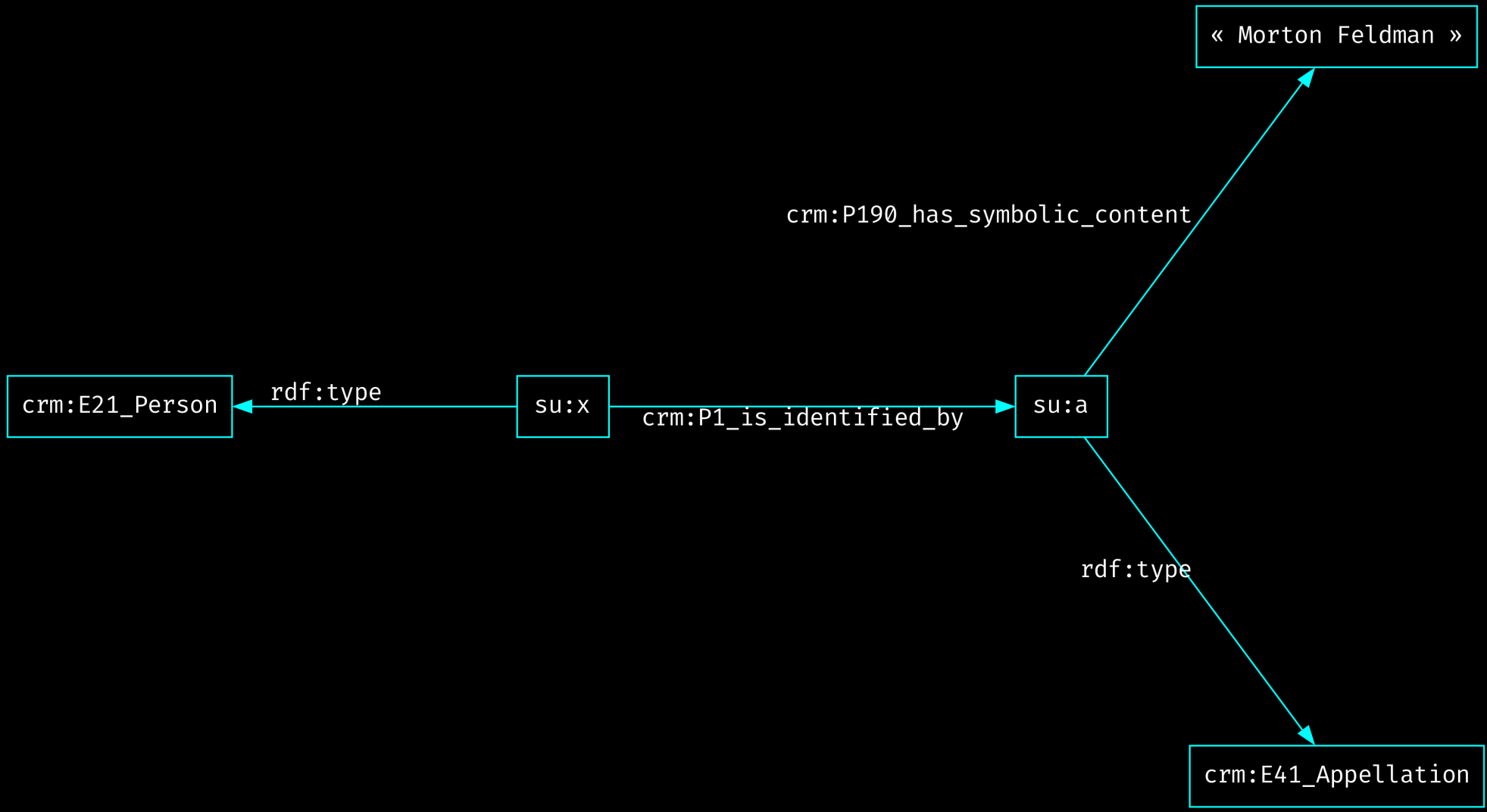
- Typer une personne :



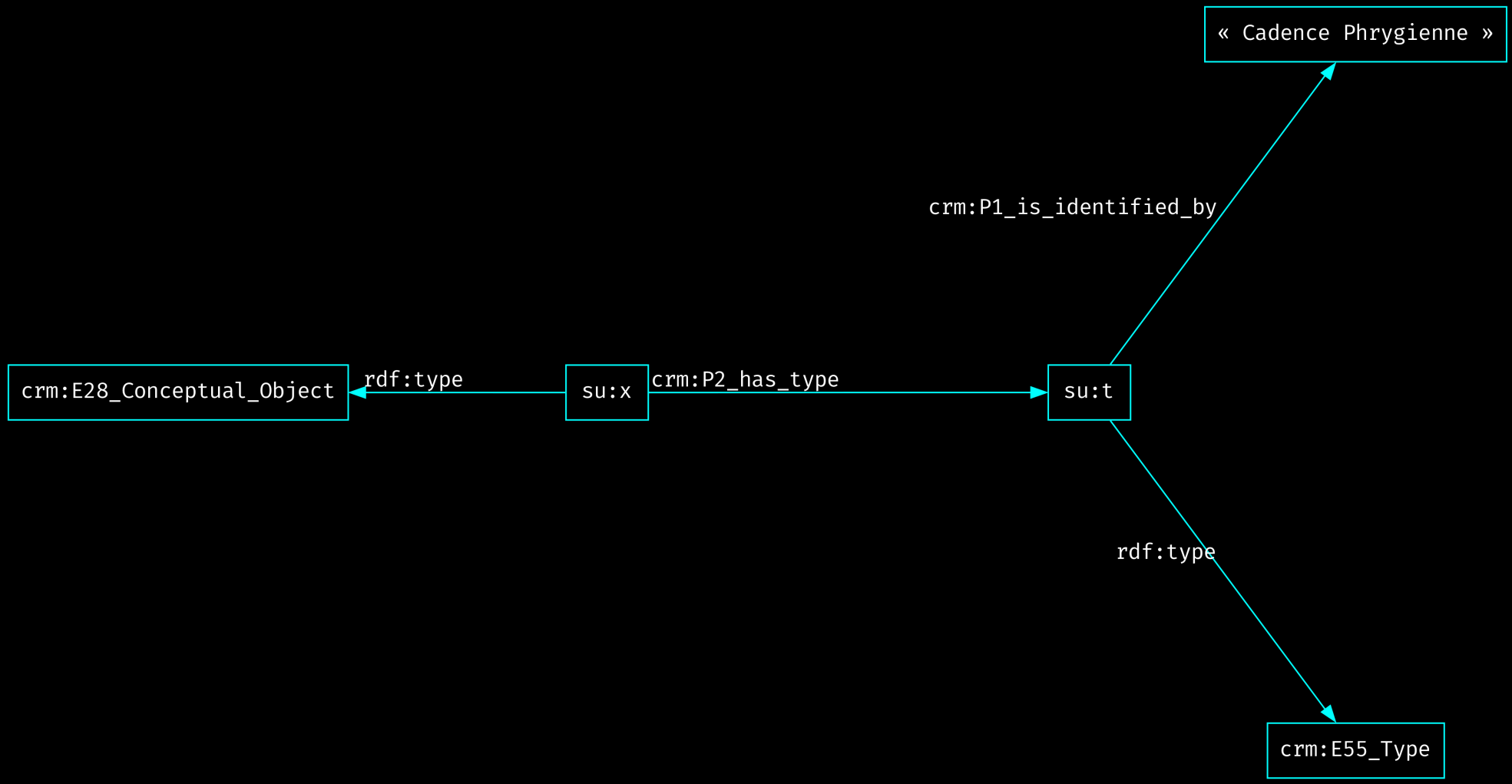
- Nommer une personne avec une simple chaîne de caractère :



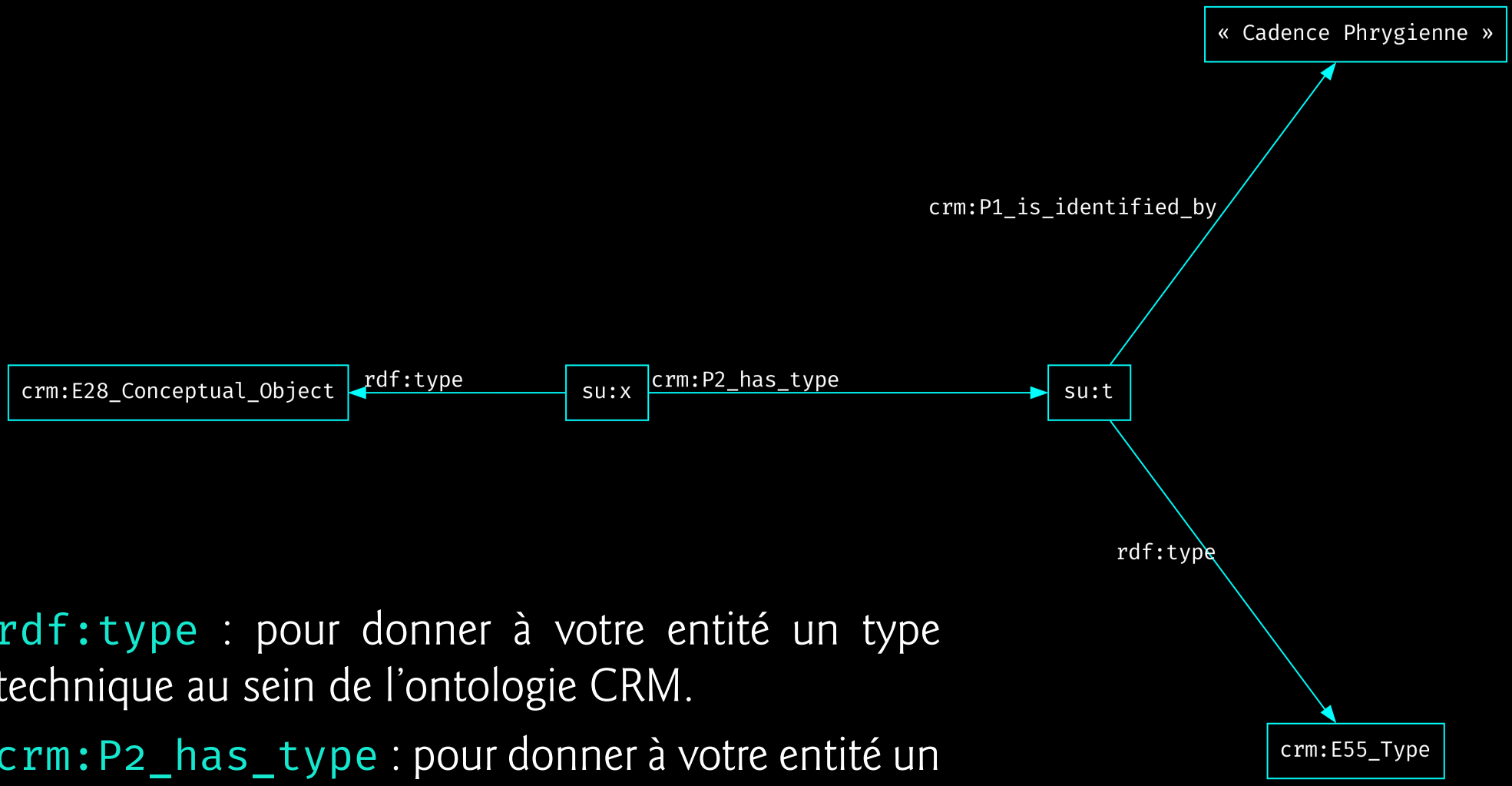
- Nommer une personne avec une entité appellation :



- Typer quelque chose :

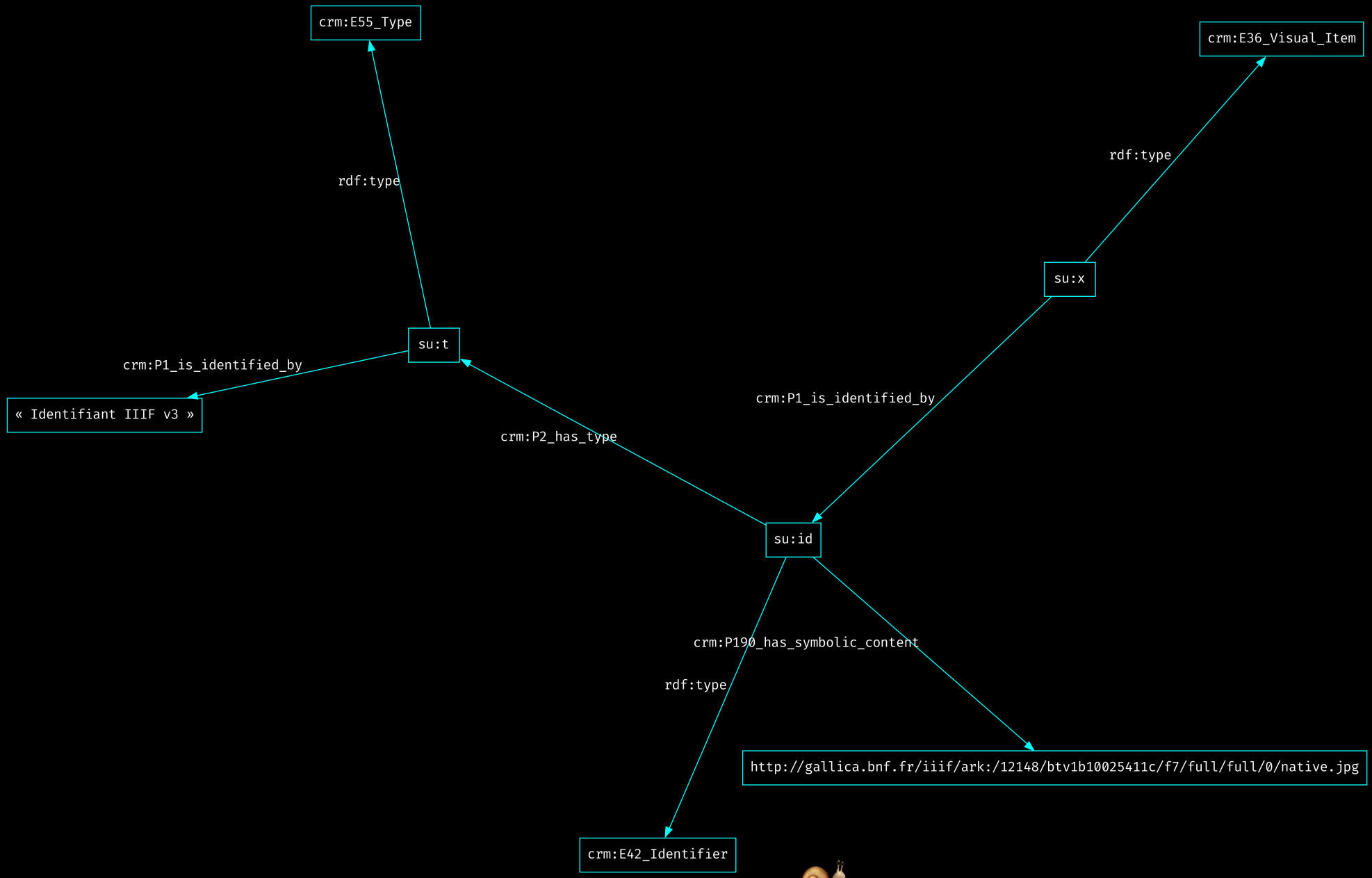


- Typer quelque chose :

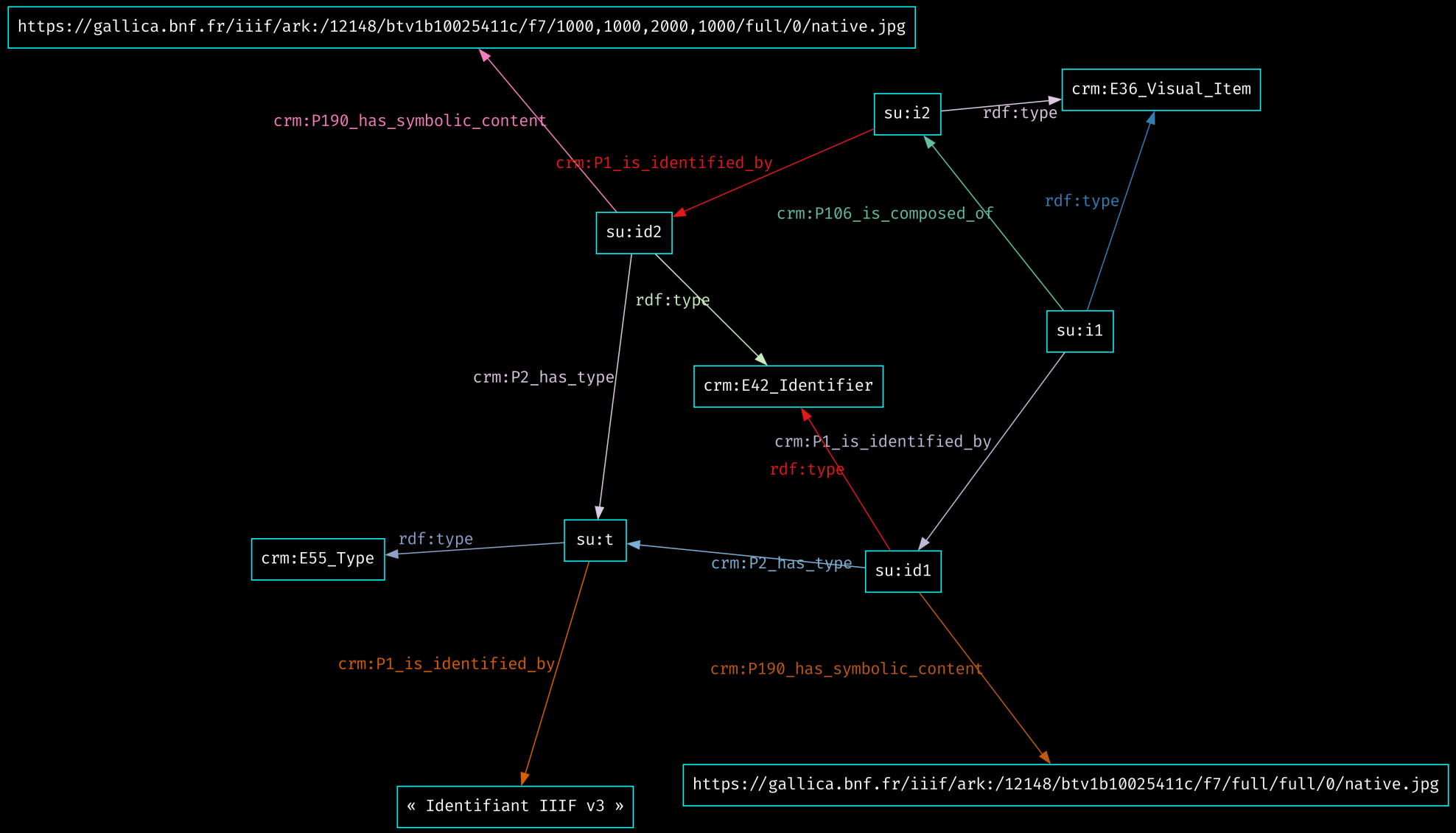


- `rdf:type` : pour donner à votre entité un type technique au sein de l'ontologie CRM.
- `crm:P2_has_type` : pour donner à votre entité un type métier (défini par vous et en rapport avec les catégories convoquées par l'activité de recherche).

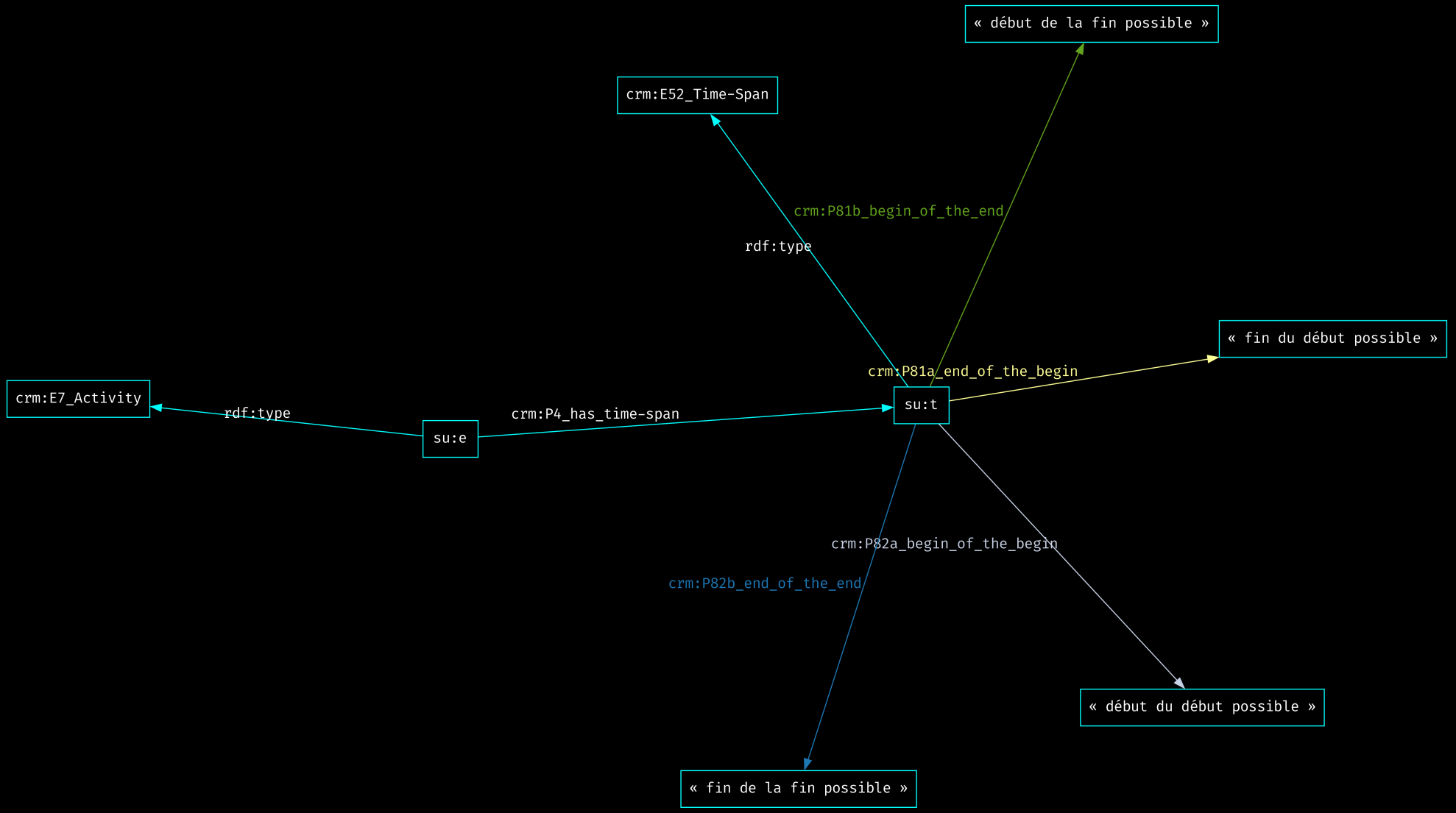
- Typer un identifiant :



- Organisation interne d’un contenu (ici, d’une image) :

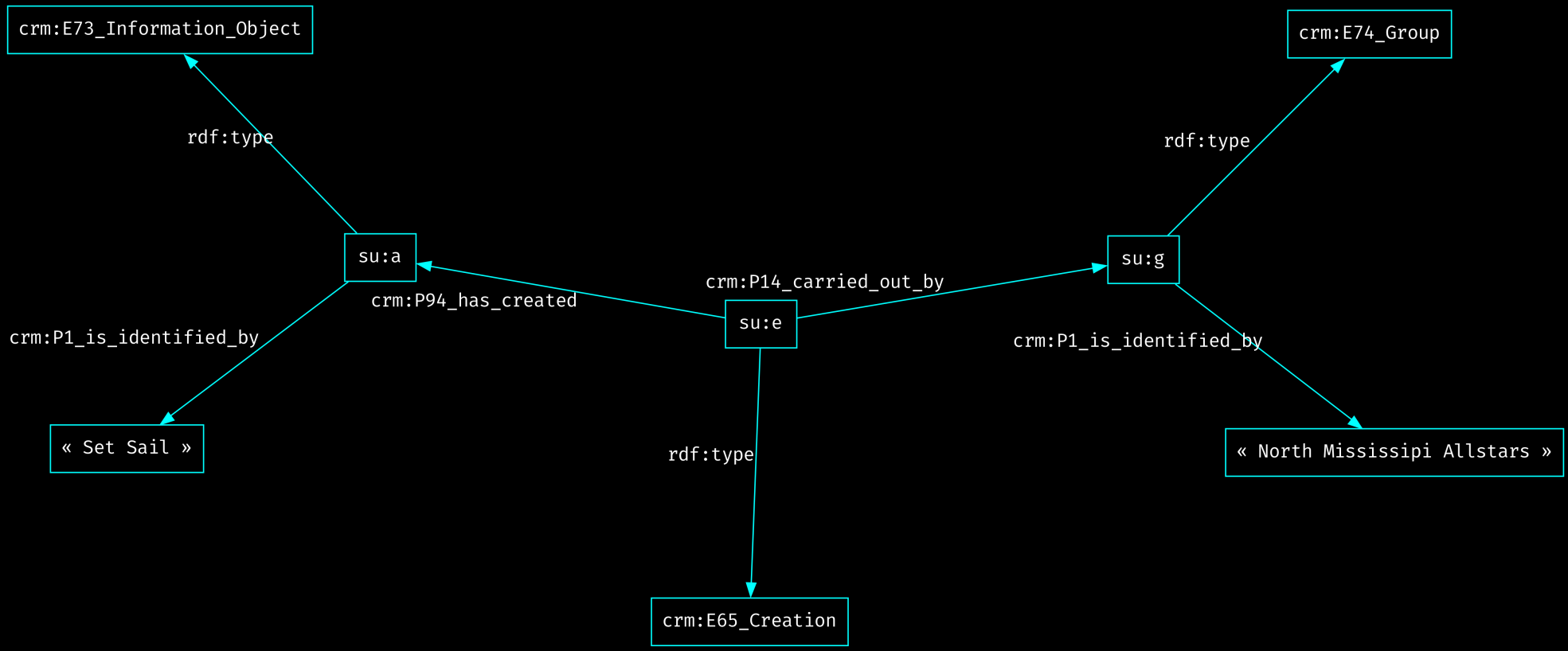


- Dater un événement :

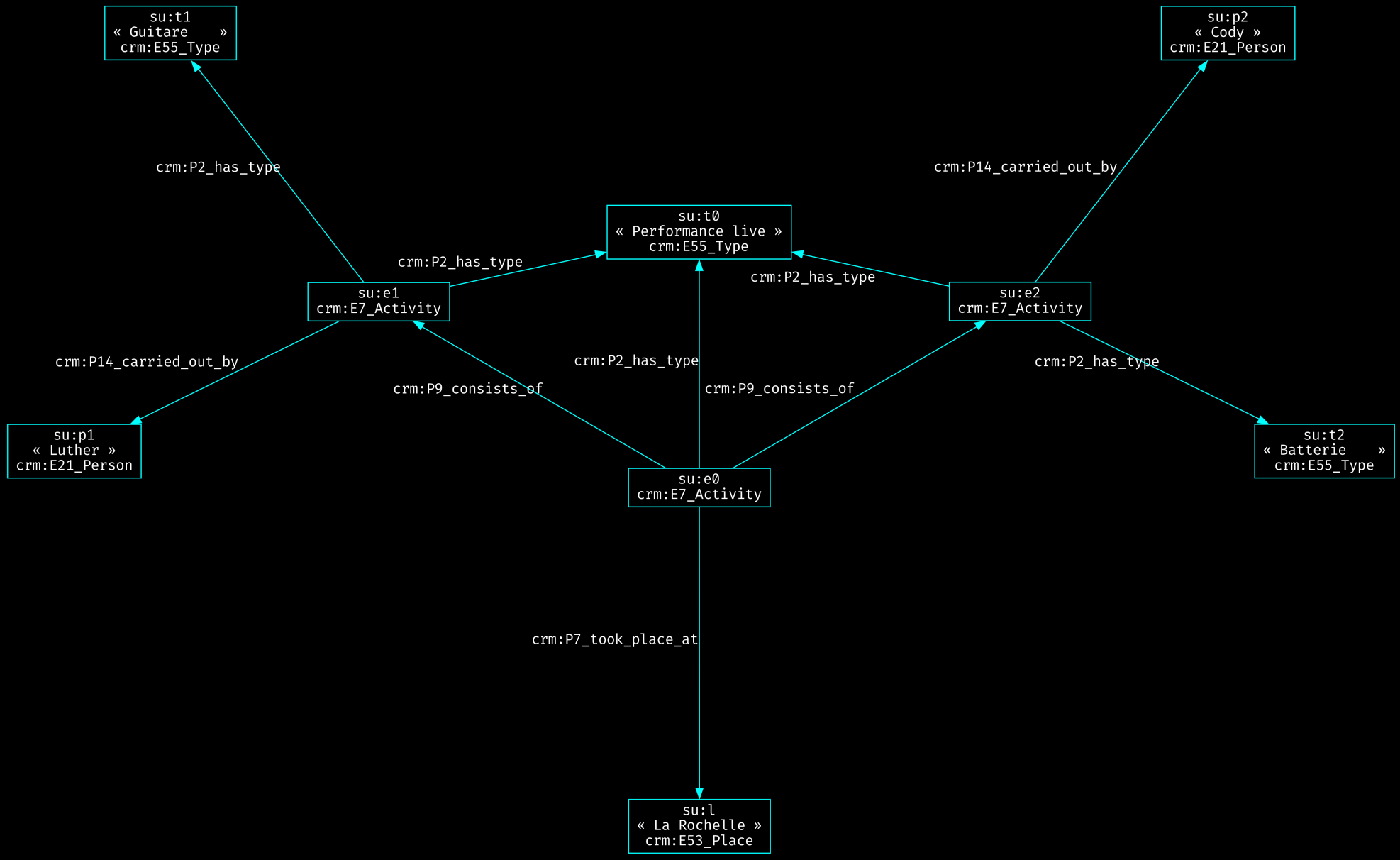


On retrouve la possibilité de définir des incertitudes aux bornes.

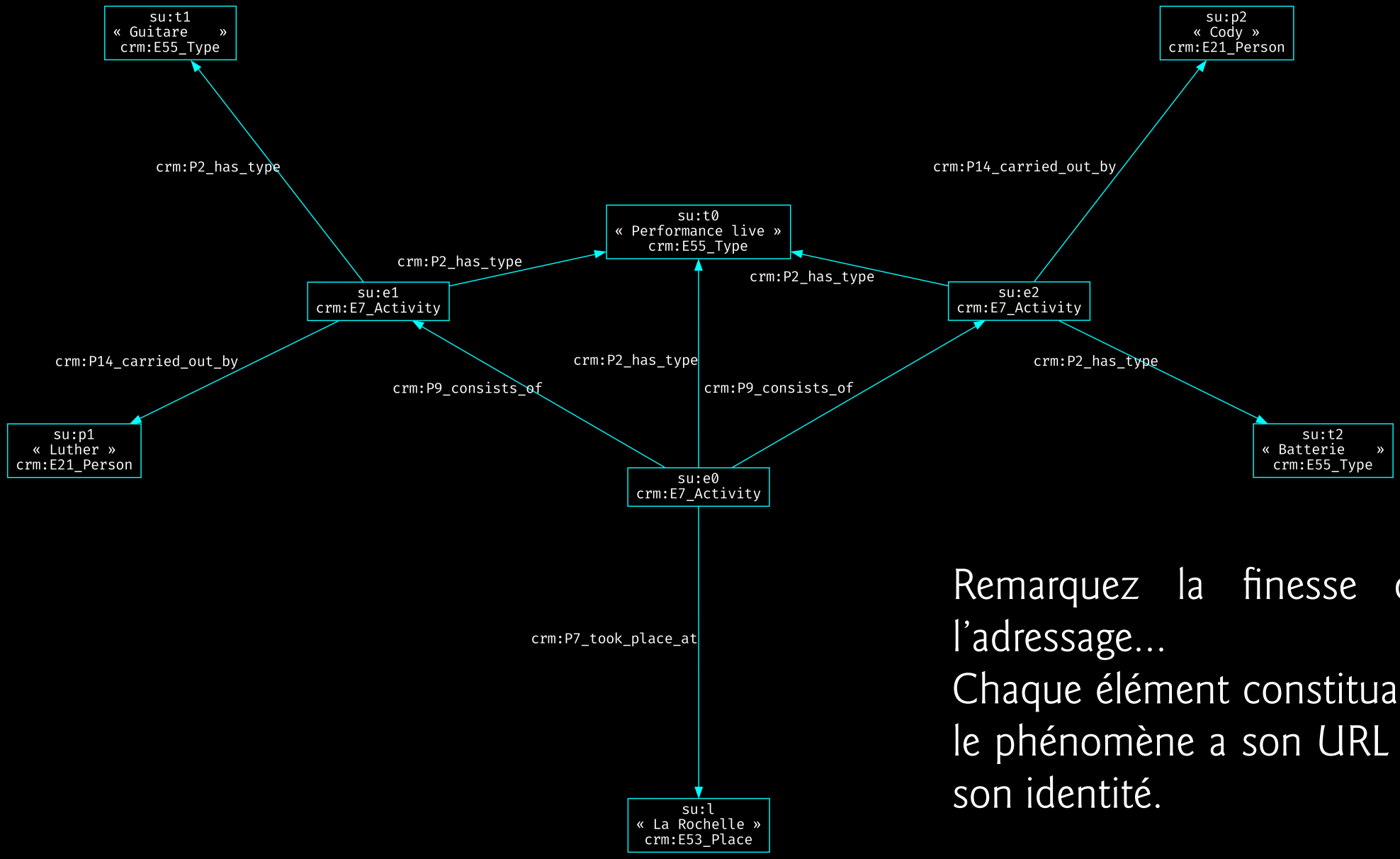
- Exprimer un événement de création simple :



- Exprimer un événement de création complexe :



- Exprimer un événement de création complexe :



Remarquez la finesse de l'adressage…
Chaque élément constituant le phénomène a son URL et son identité.

- Exprimer une annotation :