Entities, Names and Surrogates

A KR Viewpoint on a Real Example

(joint work with Michel Leclère)

Agenda

- Identification problems in CS
  - In databases
  - On the web
  - Features concerning id pbs
- Current solutions
  - Removing the problem
  - Similarity measures
  - Logical approach
- Identification pbs in digital libraries
  - Authority linking
  - Current approaches
  - Knowledge approach

Identification problems in databases

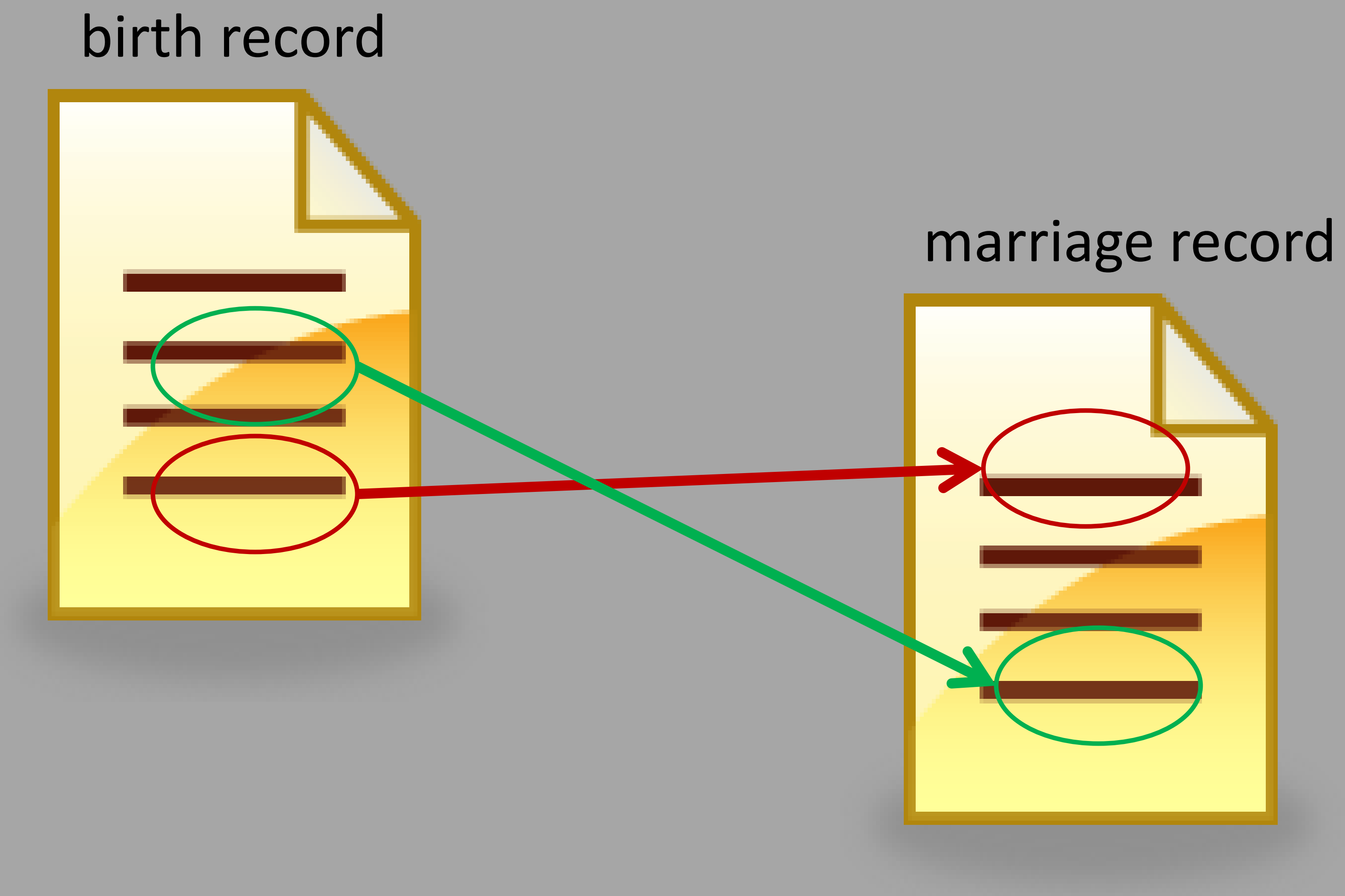
Record linkage

> “The term record linkage has been used to indicate the bringing together of two or more separately recorded pieces of information concerning a particular individual or family.”

H. Newcombe, J. Kennedy, S. Axford, and A. James. Automatic linkage of vital records. Science, 1959.

Study on family fertility in relation to the presence of hereditary disease

Linking birth records to marriage records



Record Linkage

Identification problems in databases

Two records must have *sufficiently comparable information* for making decisions about whether the records represent the same entity

Difficulties

Come from the *unreliability of the identifying information* contained in records which concern the same entity

Solution

When two records have missing or contradictory information, then the records can only be correctly matched if *additional information* is obtained

Identification problems in databases

Entity resolution, **Reference reconciliation:** which records represent the same entity, identifying multiple refs to the same object and distinguishing them from mentions of different objects

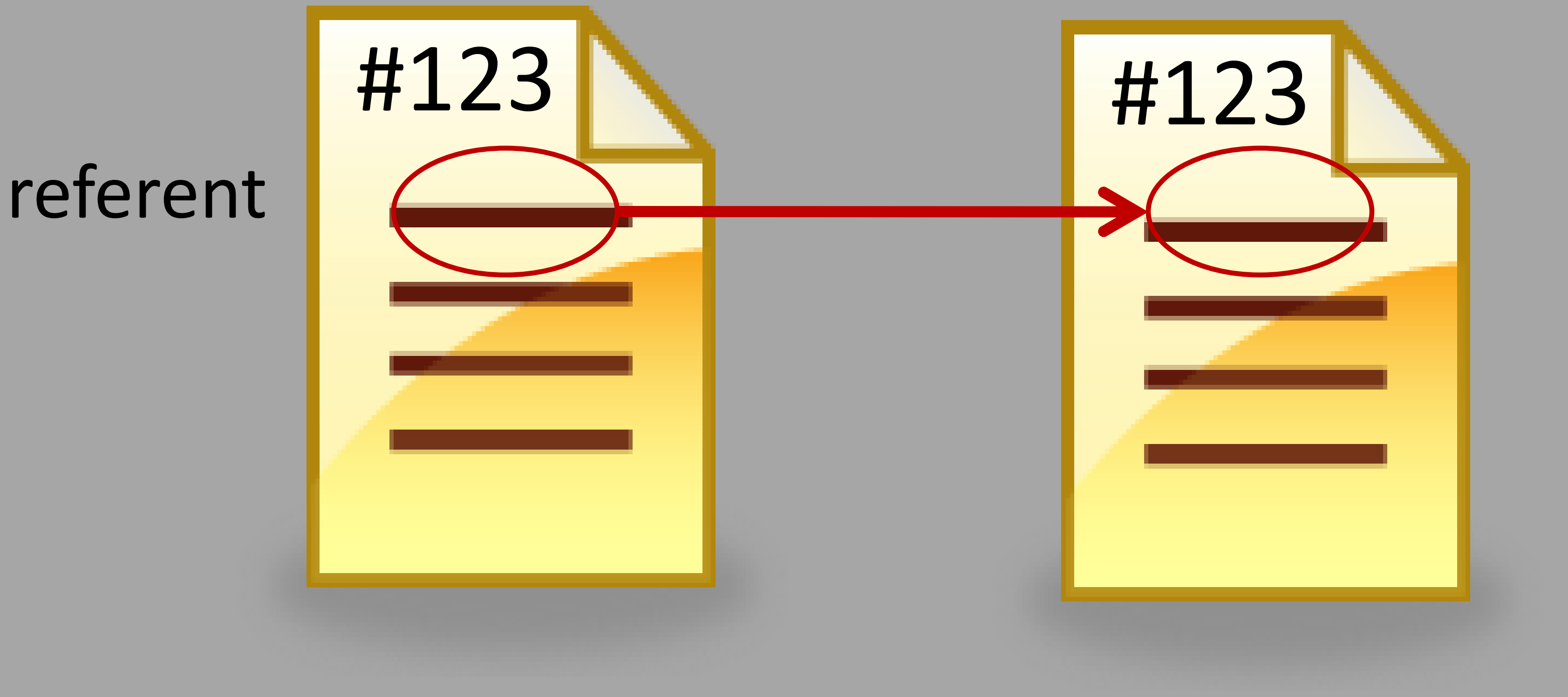
de-duplication, duplicate data (object) is deleted

Classical problem: creation of mailing lists

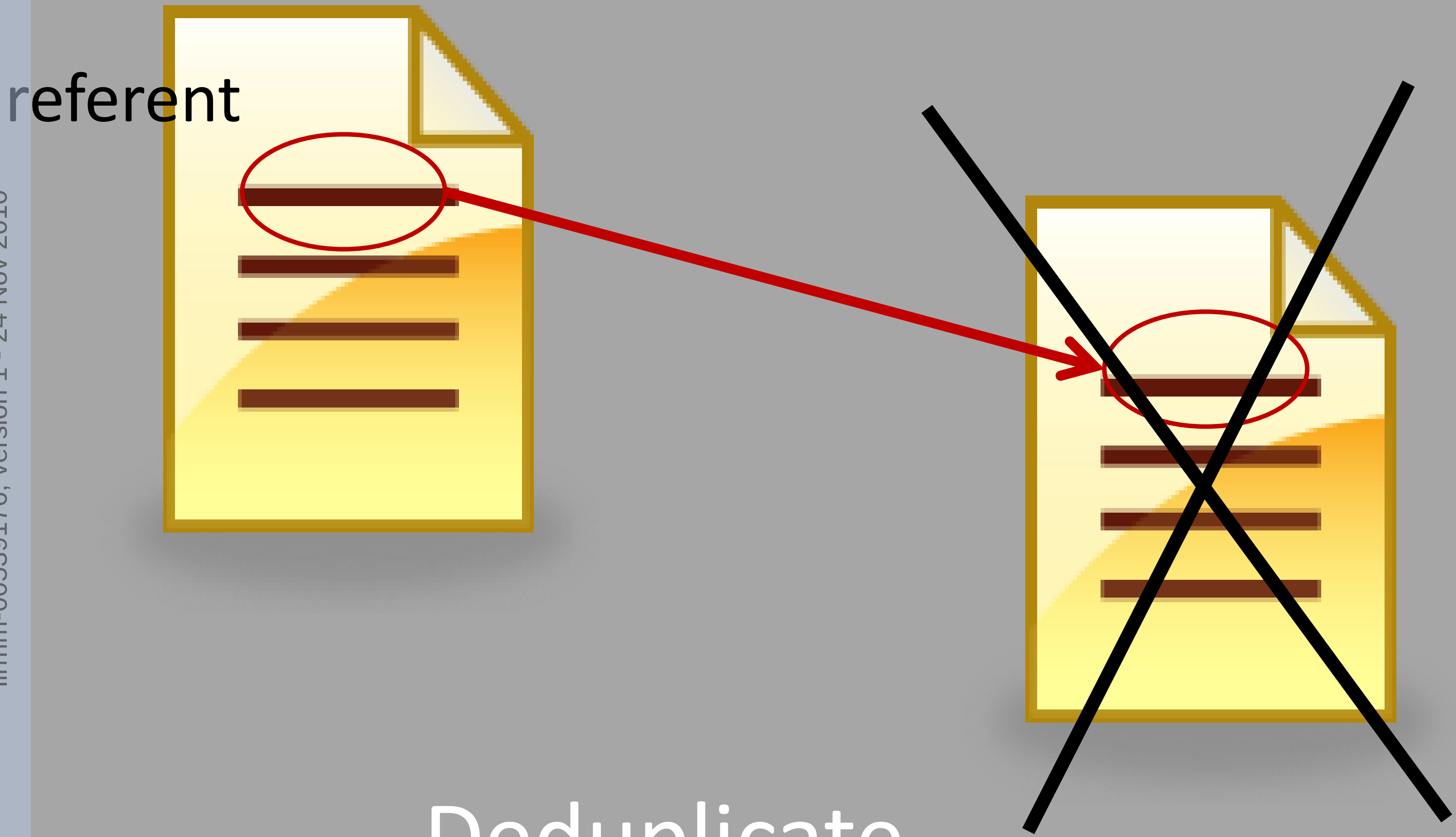
merging records judged to represent the same world entity

record linkage, linking records through refs to same world entities

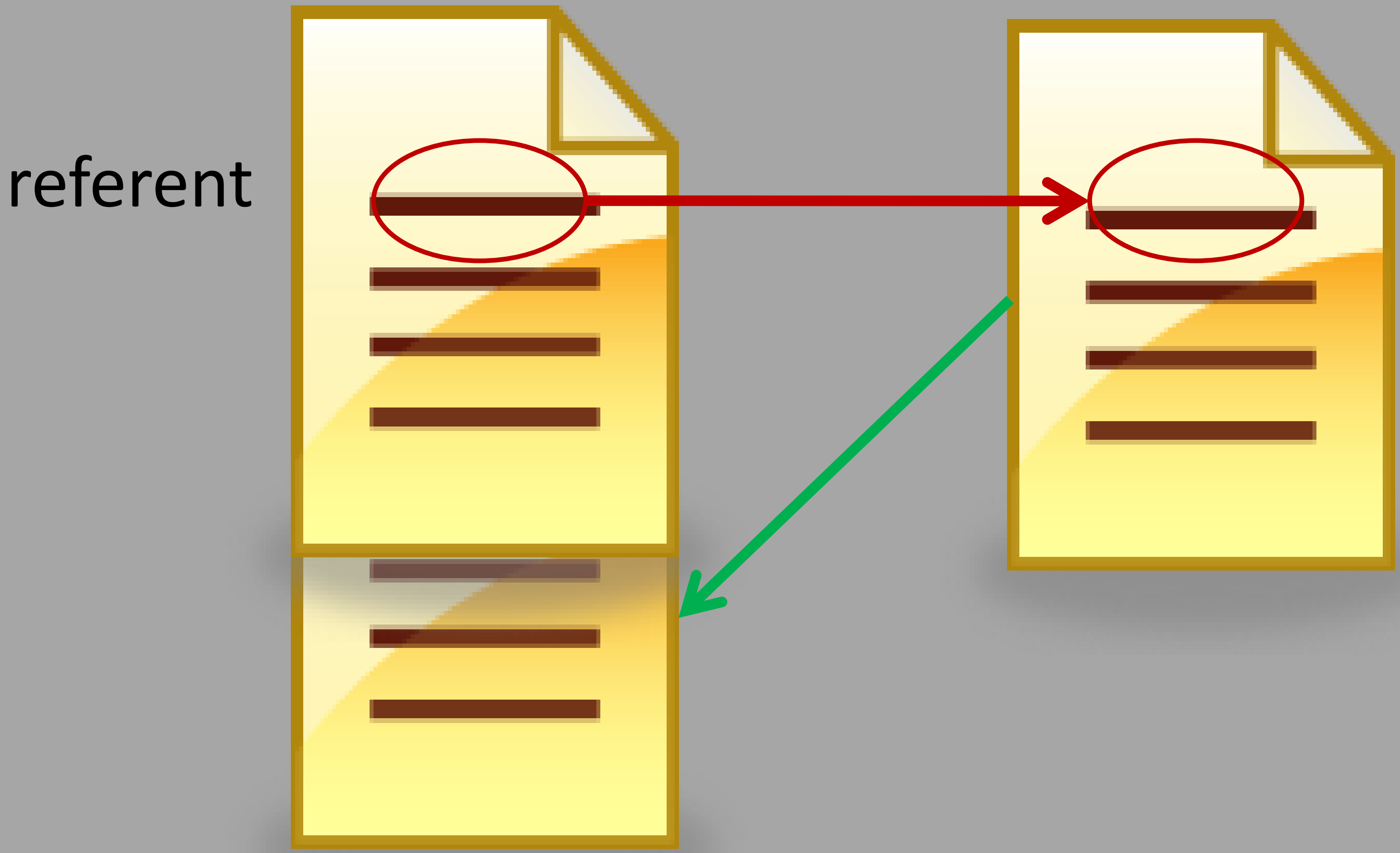
Object identification: To single out, to distinguish, to recognize



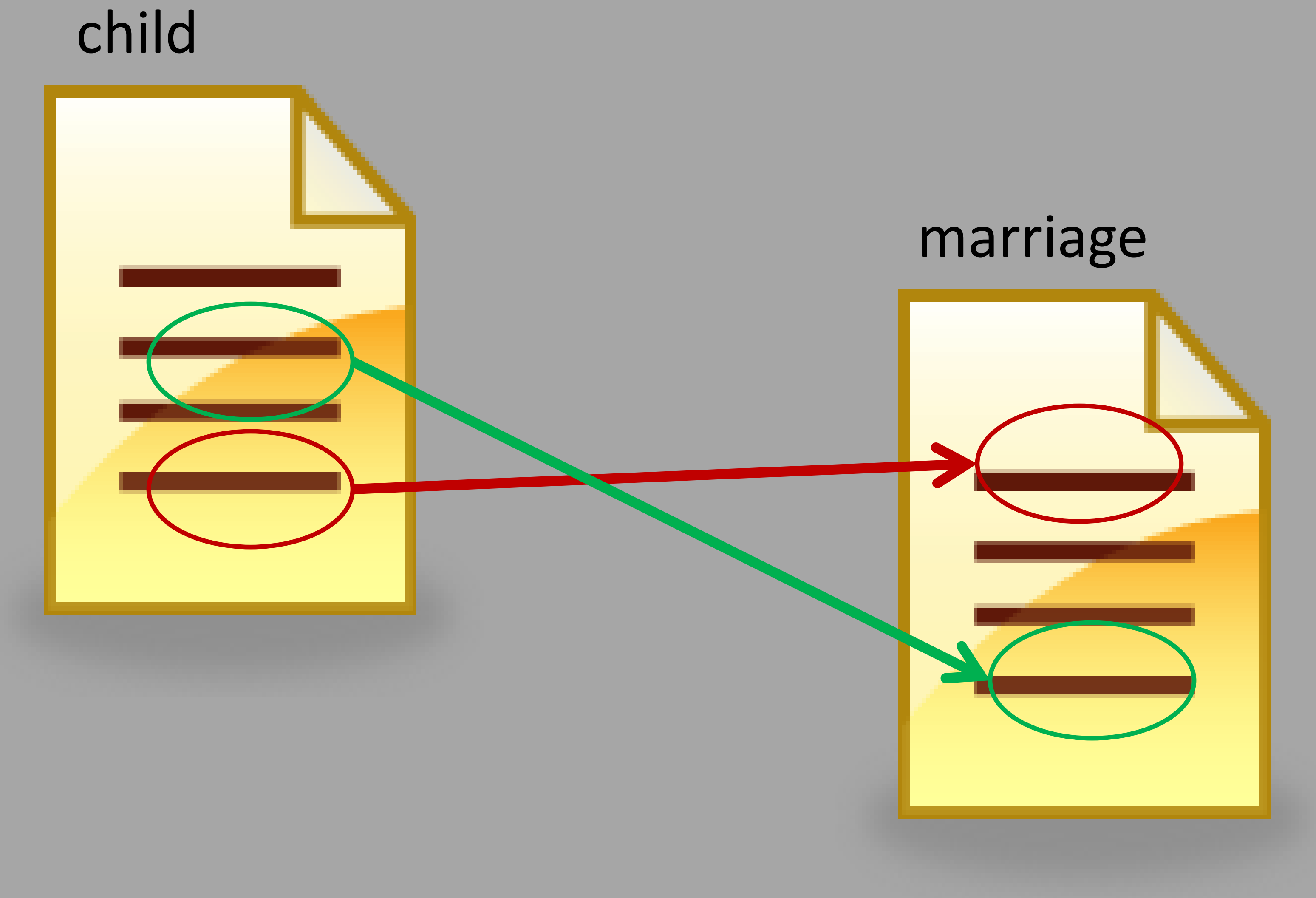
Reference Reconciliation



Deduplicate



Merge/Purge



Record Linkage

Identification problems on the web

Uniform Resource Identifiers = global identifiers of web ressources, a cornerstone of Web architecture, providing identification that is common across the Web.

The global scope of URIs promotes large-scale "network effects": a resource with an associated URI allows another party to create a link to it, make or refute assertions about it, retrieve or cache a representation of it, include all or part of it by reference into another representation, annotate it, …

Identification problems on the web

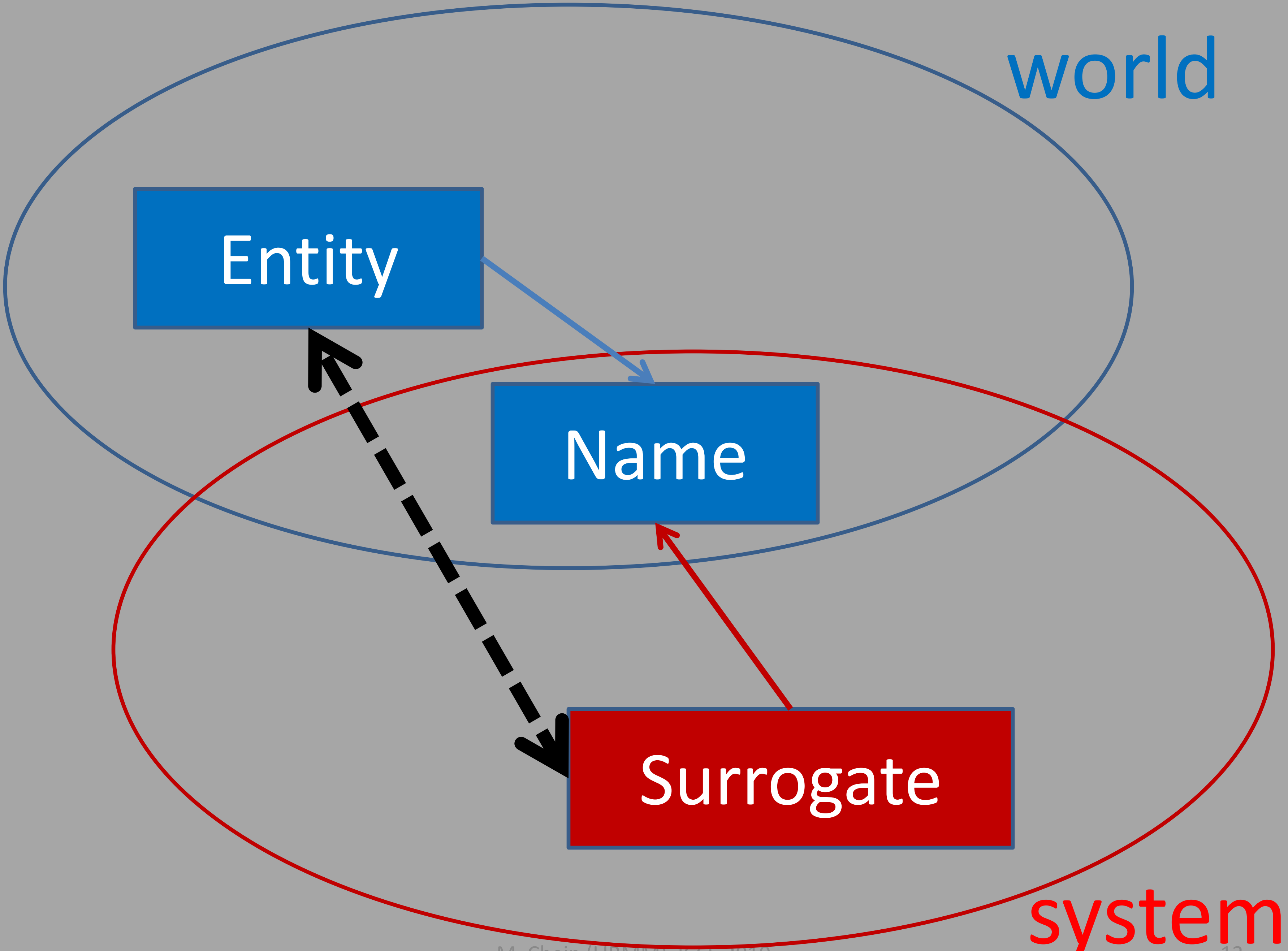
Search the web centered *on **individuals***

e.g., the various facts concerning a person if brought together form an extensively documented history of his life …

The knowledge represented in the web is gathered and entities are identified by a multitude of persons and processes for many different purposes, from many different sources.

⇒ ***inconsistencies*** occur within and between the data gathered by different processes.

"Identifying equivalent entities is a serious business"


world
Entity
Name
Surrogate
system

Collective Entity Resolution

(people,organization)
Is (the person) A within the organization B the same as A (or A') within C?
(people, people, document)
is coauthors A and B of C the same as A and B of D?"

Paul and Jane are co-authors , Jane is identified
Paul = Paul1 or Paul2?
Paul and Mary are co-authors, Mary is identified
furthermore Paul1 and Jane and Mary are co-authors
then it is possible that Paul is Paul1 in the three cases

Identification problems Features

- Natural language pbs (e.g., word sense disambiguation, generating referring expressions, web crawling) versus more or less controlled and structured information
- References may have different amount of information associated with them (e.g., an authority-person in a digital library is not a biographical record)
- Quality and amount of data
- Distinction « primary information » and « derived information » computed on a network
- Metadata (more or less structured, semantic web, digital libraries)
- PIM –personal information management: examines data from several heterogeneous sources on the desktop (files, mails, …), extract instances of different classes: Person, Message, Paper, … and relationships between instances: sender,author, …

Identification problems Wide Variety

”There is no single paradigmatic author name disambiguation task – each bibliographic database, each digital library, and each collection of publications has its own unique set of problems and issues.”

Databases differ in size, data quality, author diversity, types of metadata, rate of growth of new items, cultural context of how the data are used …

“For certain purposes (e.g., awarding the Nobel Prize to the author of a breakthrough), it may be very important to achieve a high accuracy of disambiguation.

For other purposes (e.g., as an aid to routine information retrieval), it may suffice to assign a high proportion of a person’s articles correctly, with little penalty occurring if some articles are missed or mis-assigned.” (N.R. Smalheiser, V.I. Torvik)

Removing pb

Unique Name Assumption

In databases

specializations of FOL: no function symbols and « … the two other fundamental specializations are the focus on finite models and the special use of constant symbols… for distinct constants c and c’, all interpretations that are considered satisfy neg(c=c’).”» (Abiteboul, Hull, Vianu)

In knowledge representation

Description Logics « The semantics maps each individual name a to an element a^I in D^I. We assume that distinct individual names denote distinct objects. Therefore, this mapping has to respect the *unique name assumption* (UNA), that is, if a, b are distinct names, then $a^I \neq b^I$.” (The Description Logic Handbook)

Removing the pb…

impossible in some situations!

It is possible to create a unique identifier for [illegible] would have only local significance to the creator. Anything attempting to gather data on that resource, from a foreign application, or with reference to another knowledge source would have to resolve it against existing references.

Current solution Classification

An entity is described by a set (or vector) of attributes

- attribute values are simple datatypes (e.g., strings or numbers)
- approximate similarity measures for each kind of attributes
- vector similarity measure
- weighted combination of the similarities

Ex. Graph of similarity scores for a PIM (Personal In formation Management) tool

Current solution Graph of Similarity Scores

Principles

Dependency graph: two kinds of node. Pair of references (r,r’) (which potentially refer to the same real-world entity) and pair of attribute values (a,a’), a of r, a’ of r’. Each node has a similarity score. Arc from (a,a’) to (r,r’) if the similarity (reconciliation) of r and r’ depends on the similarity (recon.) of (r,r’)

Propagation of the similarity scores from node to node in the graph

Enriching the references after merging r and r’ all the attributes of r can also be considered as attributes of r’

Enforcing constraints e.g. the authors of the same paper are distinct from each other (some node are marked *non_merge*)

Current solution Logical approach

Data sources S1 and S2 conforming to the same schema

Reconcile(a,b) iff a and b refer to the same world entity

¬Reconcile(a,b) iff they don't

A method is ***complete*** iff for each pair (a,b), a in S1, b in S2 Reconcile(a,b) or ¬Reconcile(a,b)

Current solution Logical approach

Rules

- S1 UNA $S1(x) \wedge S1(y) \wedge (x \neq y) \rightarrow \neg Reconcile(x,y)$
- $S1(x) \wedge S2(y) \wedge Reconcile(x,y) \wedge S1(z) \rightarrow \neg Reconcile(z,y)$
- LUNA for the relation R: $R(z,x) \wedge R(z,y) \wedge (x \neq y) \rightarrow \neg Reconcile(x,y)$
  $R(x,z) \wedge R(y,z) \wedge (x \neq y) \rightarrow \neg Reconcile(x,y)$
- Disjunction of classes C and D: $C(x) \wedge D(y) \rightarrow \neg Reconcile(x,y)$
- Functionality of R: $Reconcile(x,y) \wedge R(x,z) \wedge R(y,w) \rightarrow Reconcile(z,w)$
- etc

Reasoning: all $Reconcile(x,y)$ and all $\neg Reconcile(x,y)$ deduced from Facts + Rules

Conceptual Graphs

"Individual markers serve as ***surrogates*** that uniquely identify the individuals that are cataloged in a conceptual system … ***names are demoted to the status of characteristics*** that are no more fundamental than weight or hair color (since one entity may have multiple names or *aliases*, and multiple entities may have the same name)."

"Since surrogates are unique only within a particular computer system, ***communication between system still depends on printable names*** with their potential ambiguities."

"There are no URLs for things that cannot be flattened out and stored on a computer disk, such as dogs, trees, and people. For such things, the surrogates serve as local substitutes within a database. But the task of matching the surrogates to the physical objects cannot be done wholly within a computer or even a network of computers. There must some sense organs –human or robotic- that can ***relate the internal identifiers to the physical world***."

(Sowa)

Digital Libraries

Digital Libraries Features

Bases of notices

Notice = metadata

Bibliographic notices: metadata associated with a document

Authority notices : metadata associated with instances of specific classes (Person, Collectivity, Geo_place, Subject, …)

Each notice has an identifier: surrogate

Relationships between bibliographic notices and authority notices (authorOf, editorOf, …) are values of attributes

The set of metadata bases can be seen as a structured graph

Digital Libraries Problems

Dynamic

- Adding notices to a base
- Merging bibliographic bases
- Making data accessible to outside

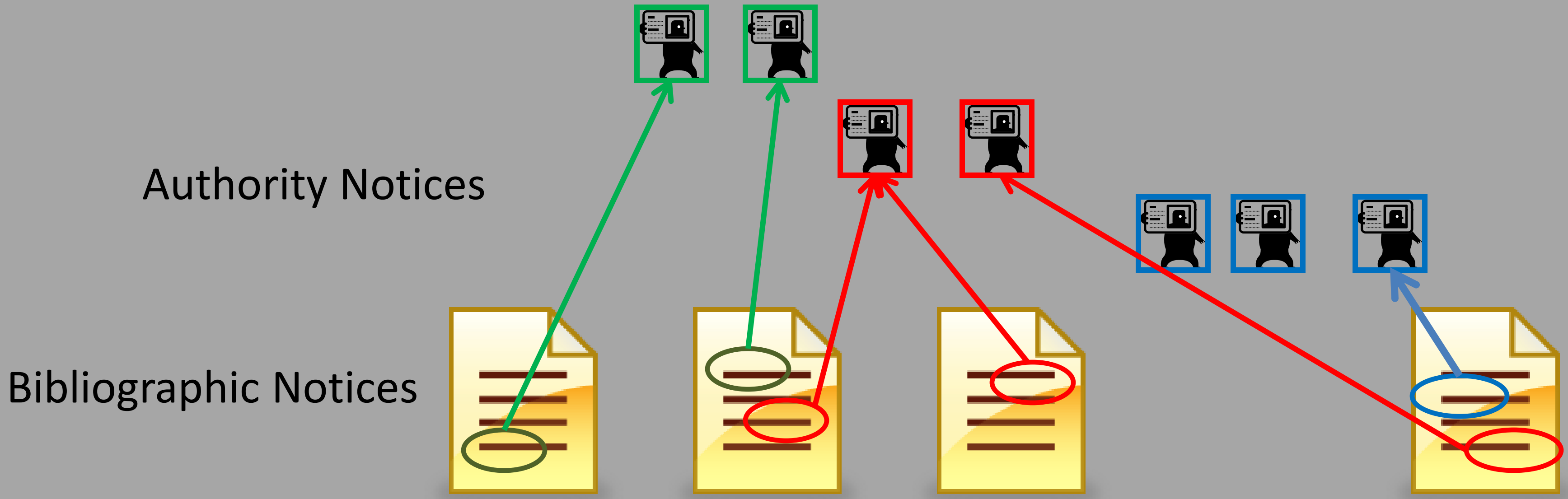
Quality of the bases of notices

- Consistency inside and between bases
- Relevance of the subject
- Maintenance of the different bases

Digital Libraries Authority Linking

Goal: building a service of authority identification

Identifying in a new notice references to an entity co-referent with an authority (person, collectivity, subject) in a bibliographic notice



Digital Libraries

Current Approaches

Manual linking

- Given a « term » and an « authority type », e.g. (« Victor Hugo », Person) search the person authority file containing « Victor Hugo » as a normalized or a rejected form
- Authority notices are returned to the user
- The user assigns the more relevant authority to the term (links the term to the identifier) or create a new authority

Automatic linking by similarity scores

- Similarity score for each attribute of the authority that is present with « term»
- Aggregation of similarity scores
- Authorities ranked by decreasing score, the first is considered if there is a gap with the second

Knowledge Approach

Exploitation of the information in the network of notices

- Representation of the database in GBKR (our CG model)
- Enrichment of the authority notices
- Identification of authorities in a new notice

Knowledge approach 1 Ontology

- FRBRoo 1.0 (2009)

  « a formal ontology intended to capture and represent the underlying semantics of bibliographic information and to facilitate the integration, mediation, and interchange of bibliographic and museum information »

  Hierarchy of classes (Work, Expression, Manifestation, …), hierarchy of relations with signature (brought_into _life (Birth, Person), has_fragment(Expression_fragment, Expression), …

- Representation in RDFS then in a GBKR vocabulary
- Features :
  - Properties and subproperties of FRBRoo represented as ordered relations
  - Distinction between concept and data
    - A title vs. a string on the cover
  - Metadata concerning the notices (date, origine, sources…)

Knowledge approach 1

Authority notice

```
001A    $0751062103:02-12-04
001B    $0751062103:02-12-04$t11:43:29.000
001D    $0751062103:02-12-04
001U    $0utf8
001X    $00
002@    $0Tp5
003@    $0XXXXXX36
010@    $S##$afre
012C    $S##$a0$b1$c0
012E    $S##$ab
019@    $S##$aFR
028A    $S#1$40y$dChristian$aBernard
037F    $S##$aDessinateur de bandes dessinées
047M    $S##$aHépatite virale C ; ça craint ! / Dr Léo Py, Christian Bernard, 2003
```

ppn: XXXXXX36 Vedette Nom de personne

Norm. Form: Bernard, Christian
Forme savante ou à valeur internationale

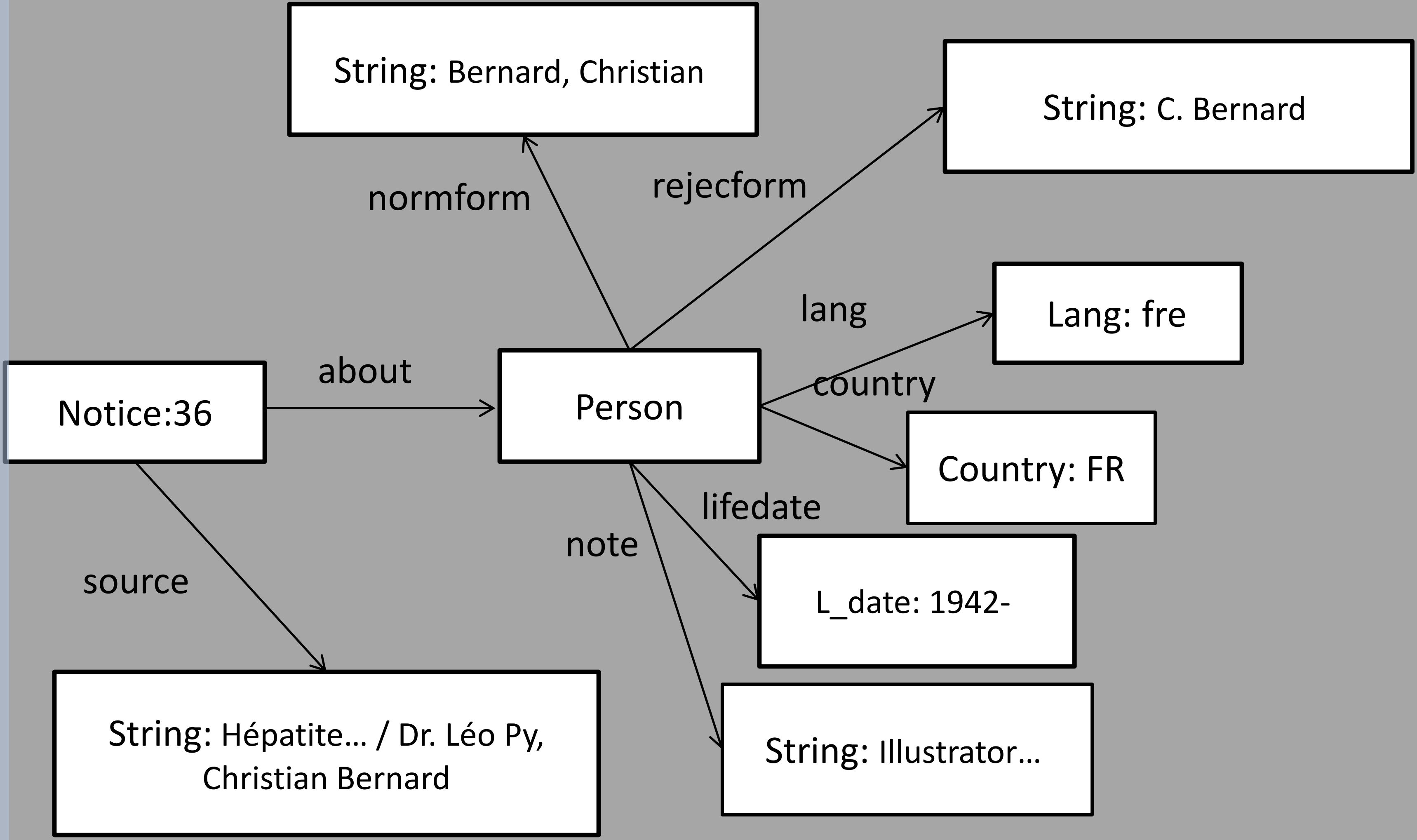
Country : France

Lang : français

Note : Dessinateur de bandes dessinées

Source : Hépatite virale C ; ça craint ! / Dr Léo Py, Christian Bernard, 2003

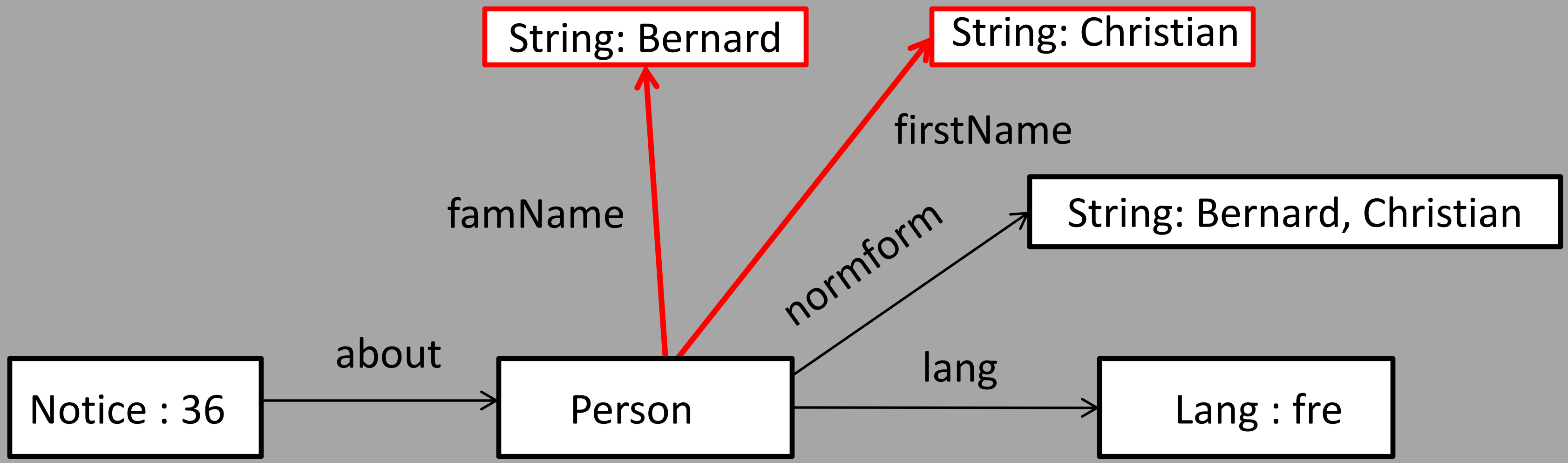
Authority notice



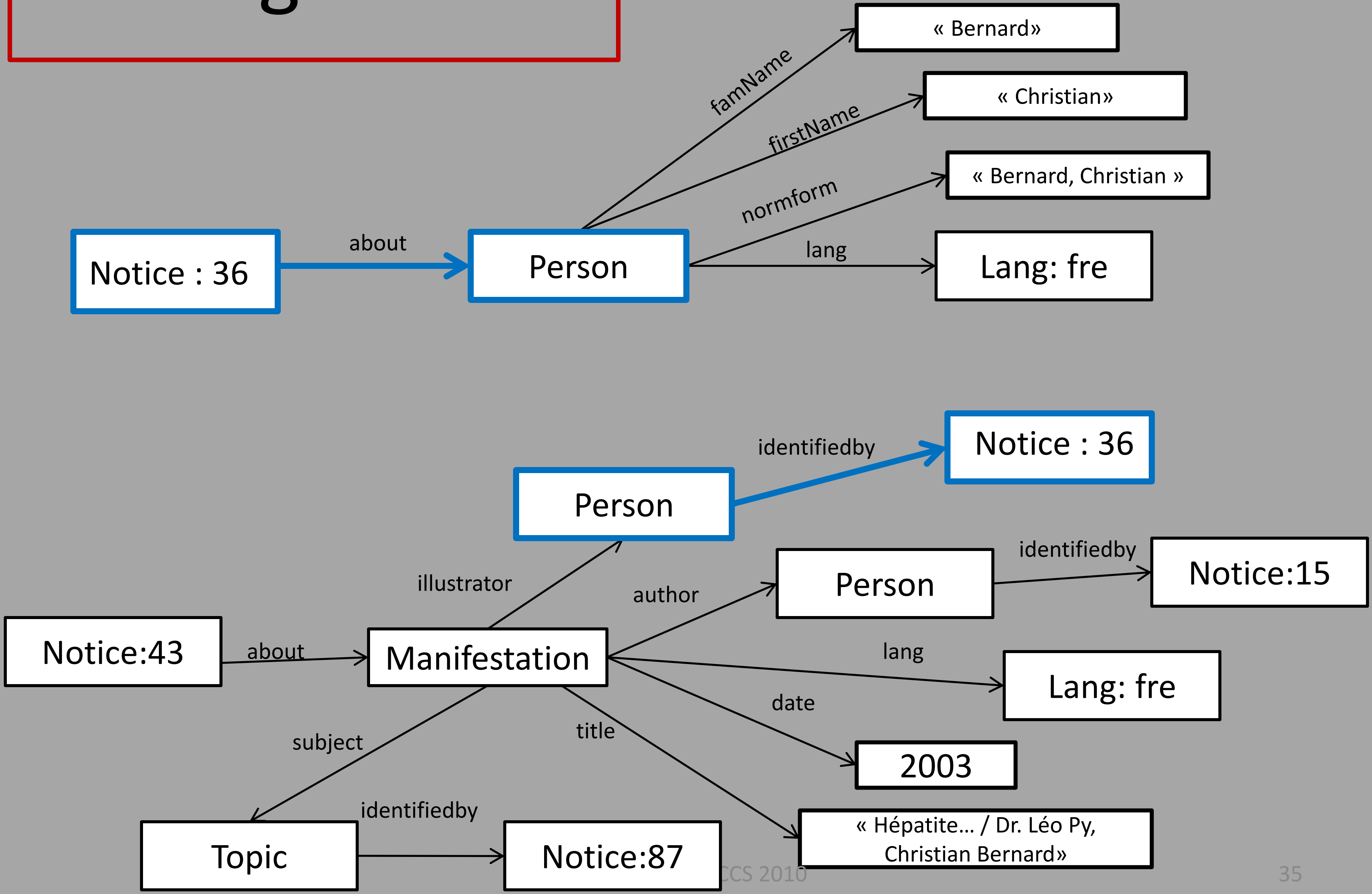
Authority Notice Enrichment

- Elicitation of knowledge in authority notices
- Exploitation of links between bibliographic and authority notices (merging of entities)
- Inference rules

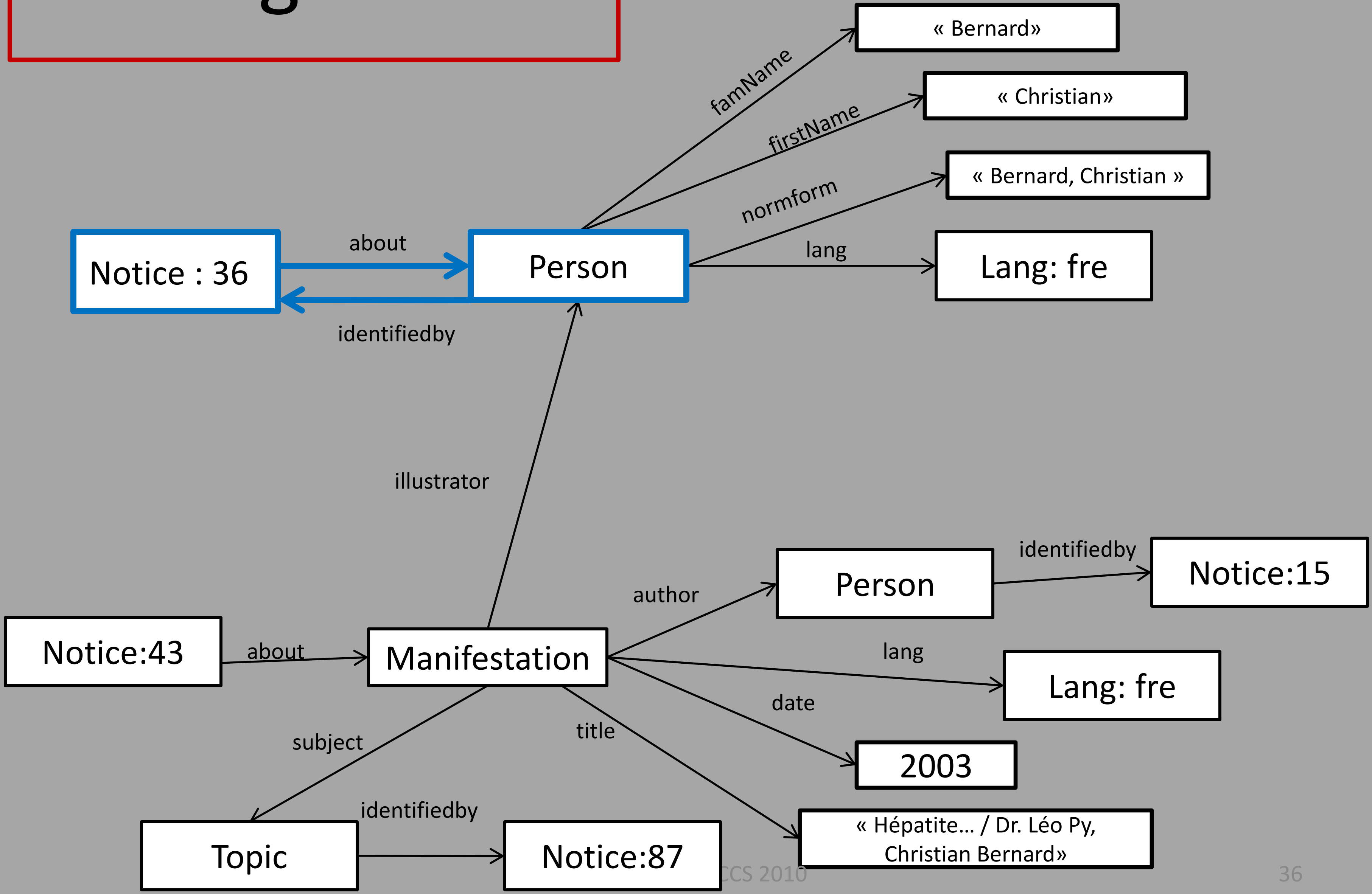
Knowledge elicitation



Linking notices

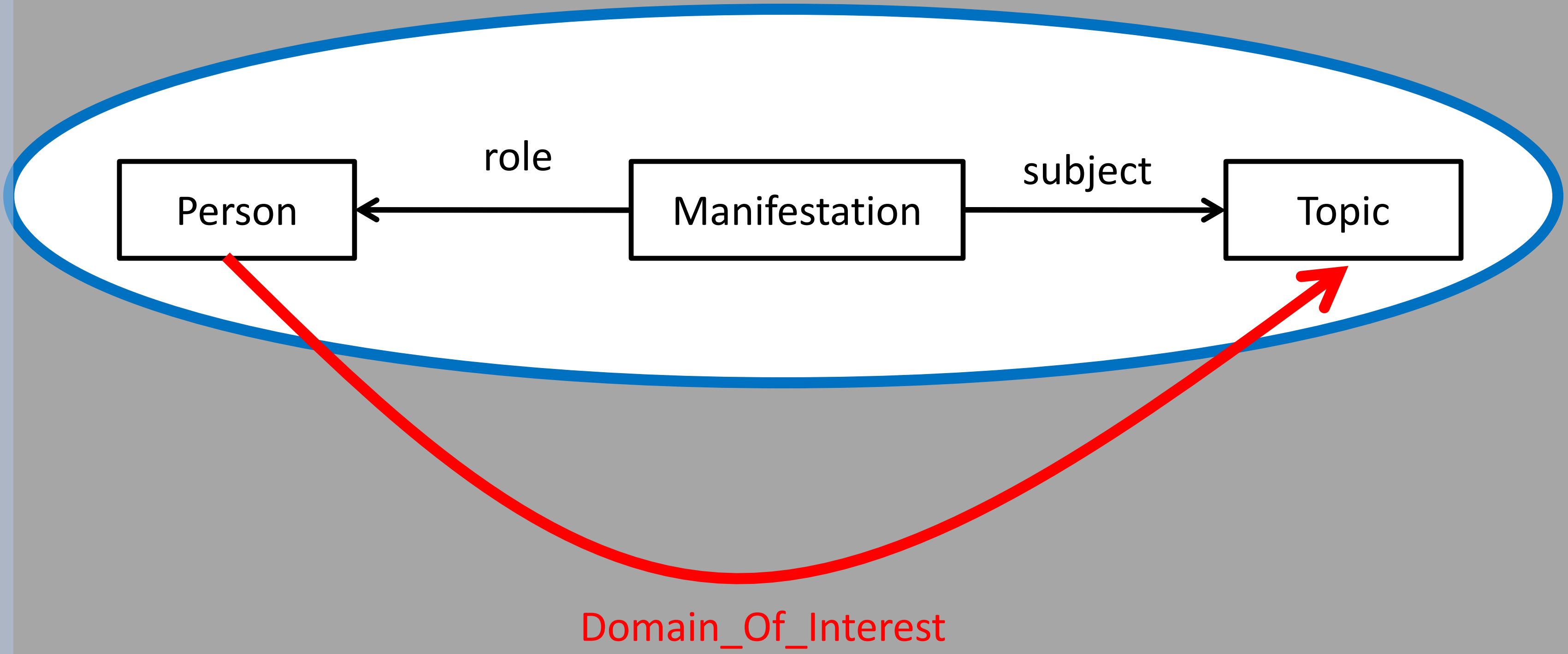


Linking notices

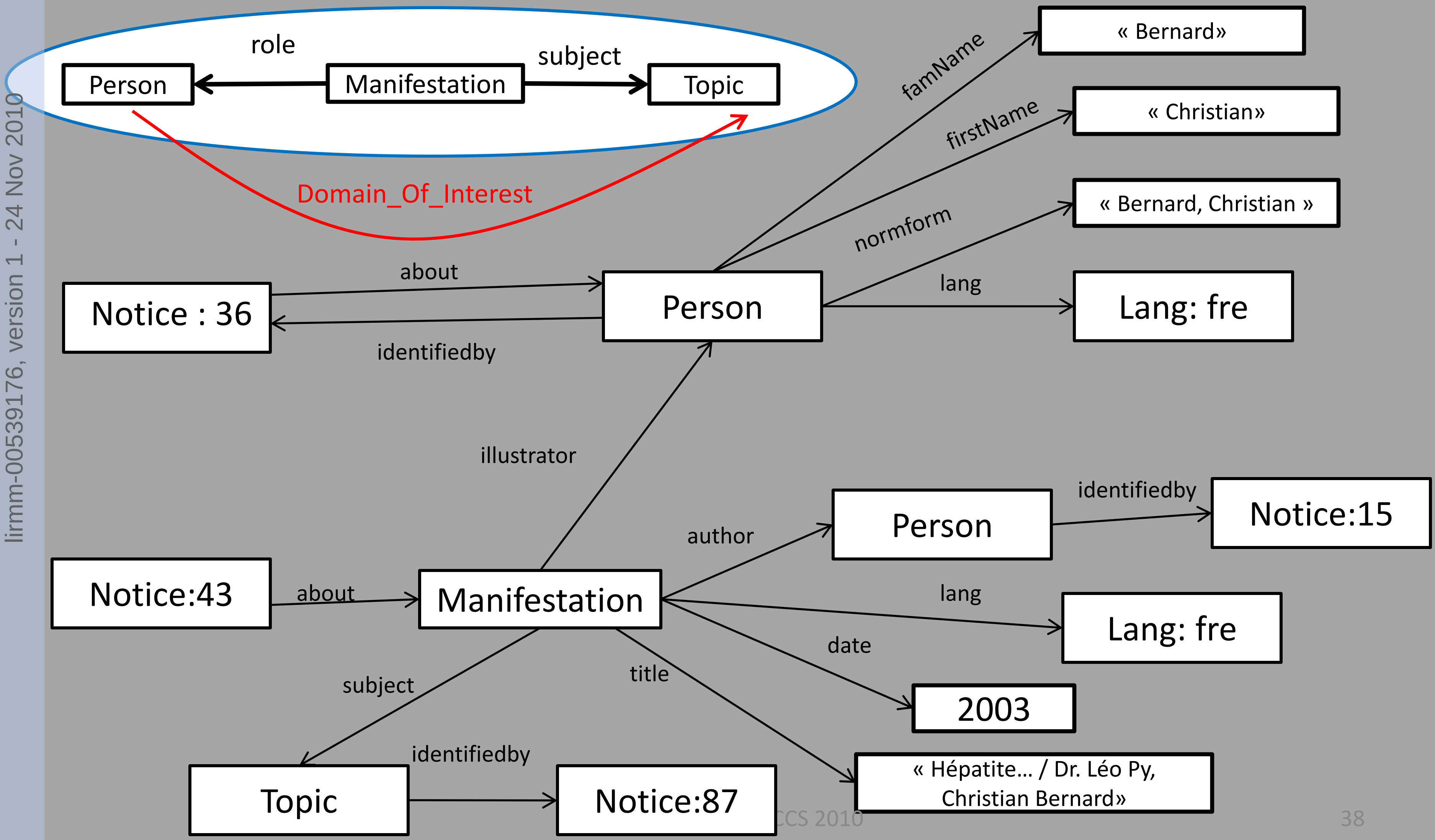


Inference rules

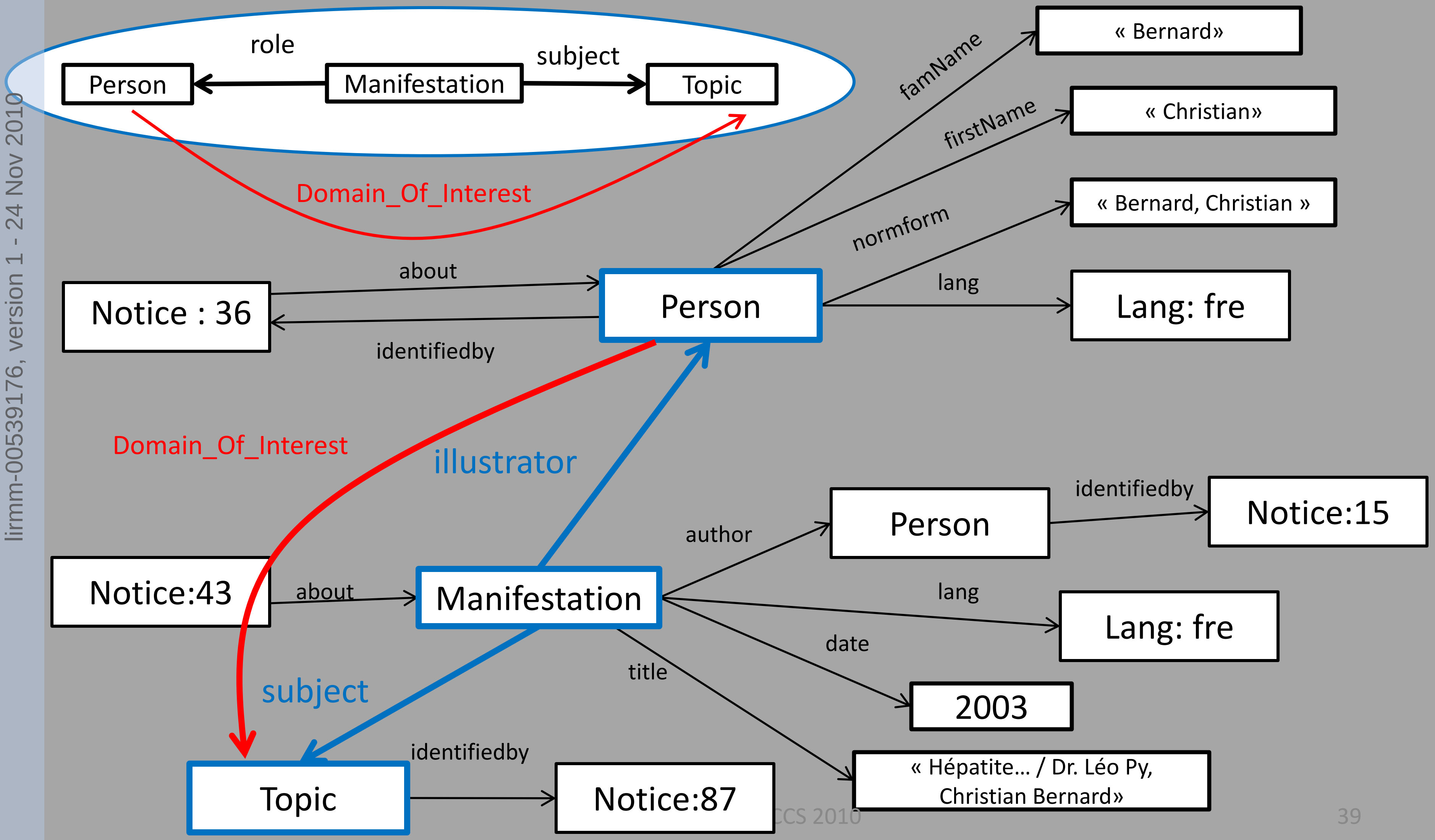
If H is present then C can be added



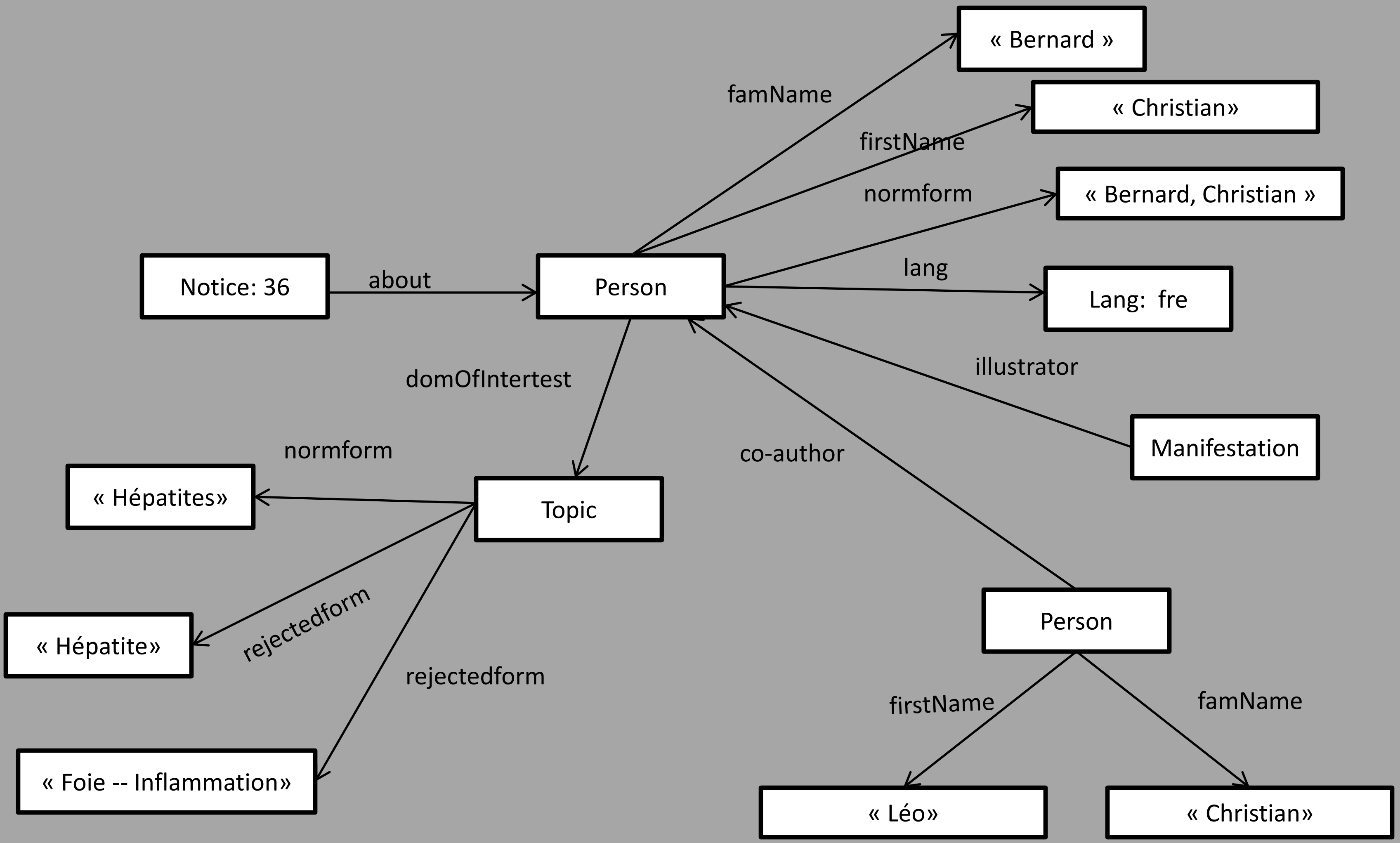
Linking notices



Linking notices



Enriched authority notice



Authority Identification

Selection pattern: necessary and optional information

Querying the enriched notice base

- Necessary information: filter
- Optional information: ranking

Collective consistency checking

Identification Selection pattern

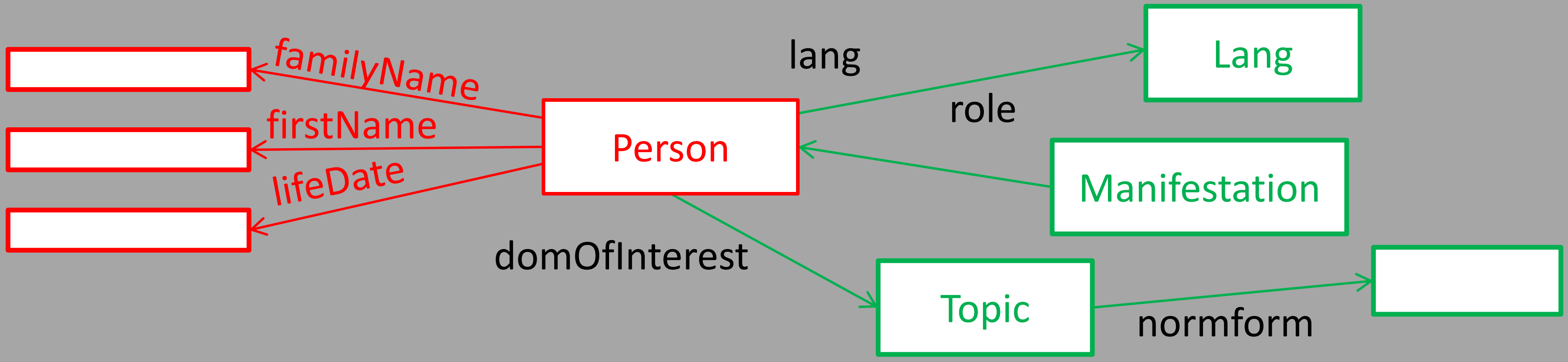
Selection pattern: a query in information retrieval

Necessary Part strong matching with enriched authority notice

Used to select a set of candidate authorities

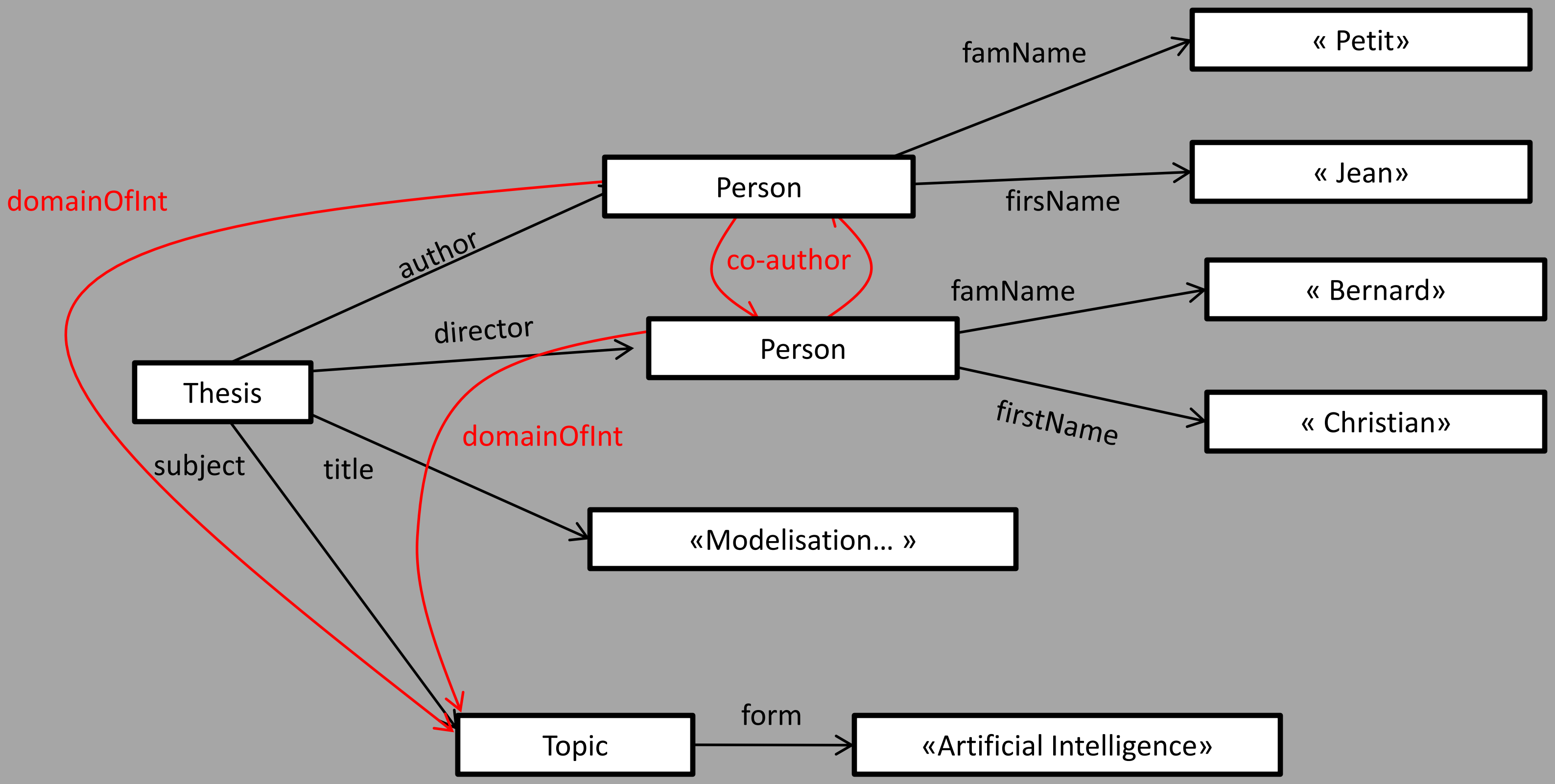
Optional Part used to rank the candidate authorities

Example : pattern for person



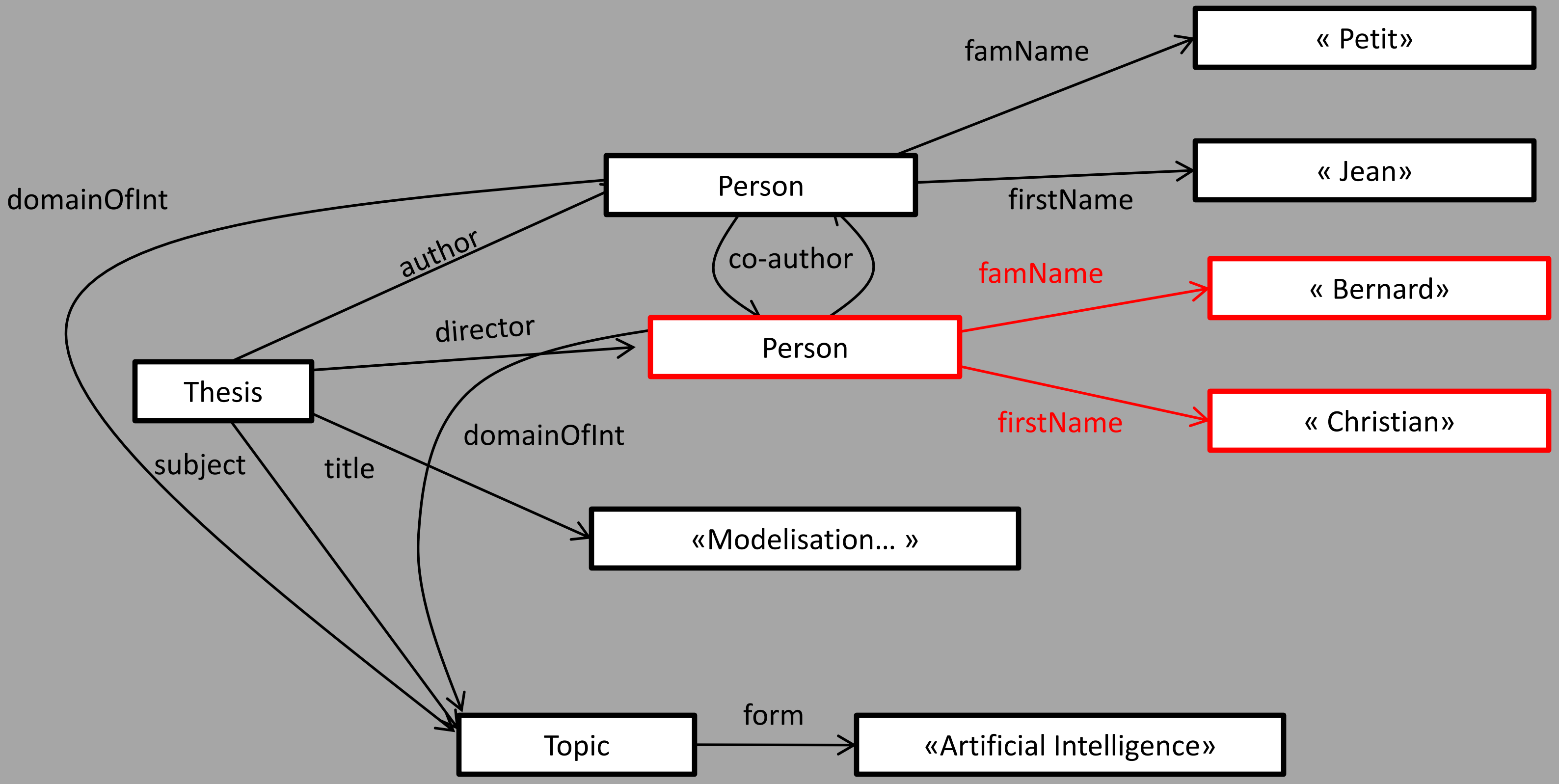
Identification A new bib notice

Represented in the language and enriched

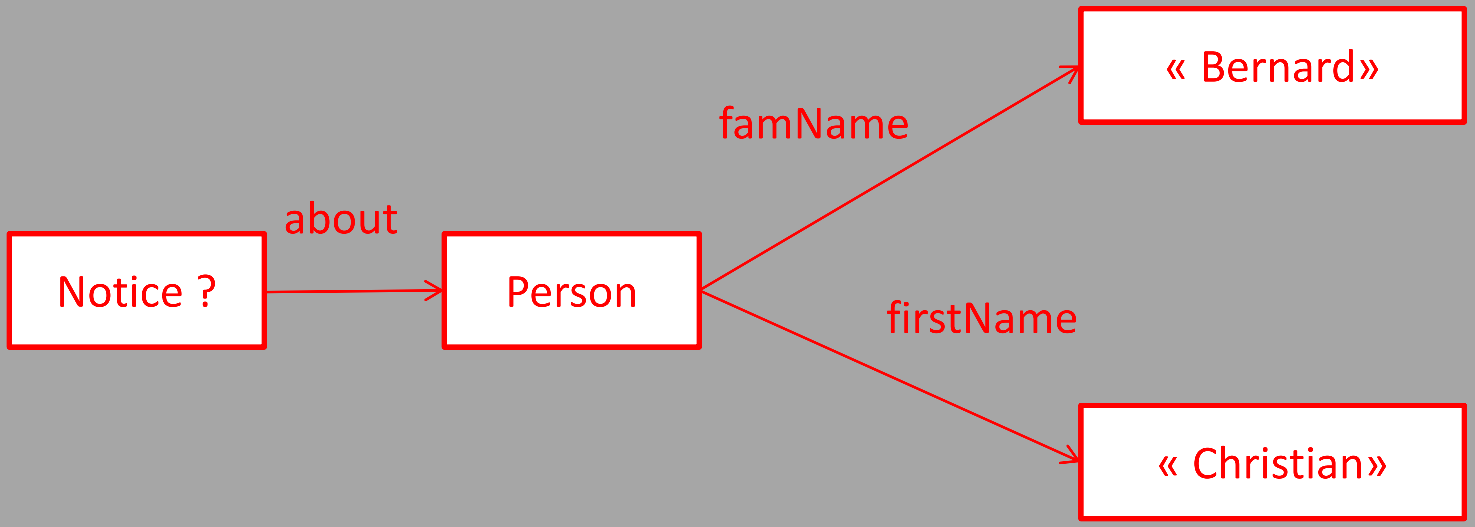


Identification Selection

Choose the selection pattern

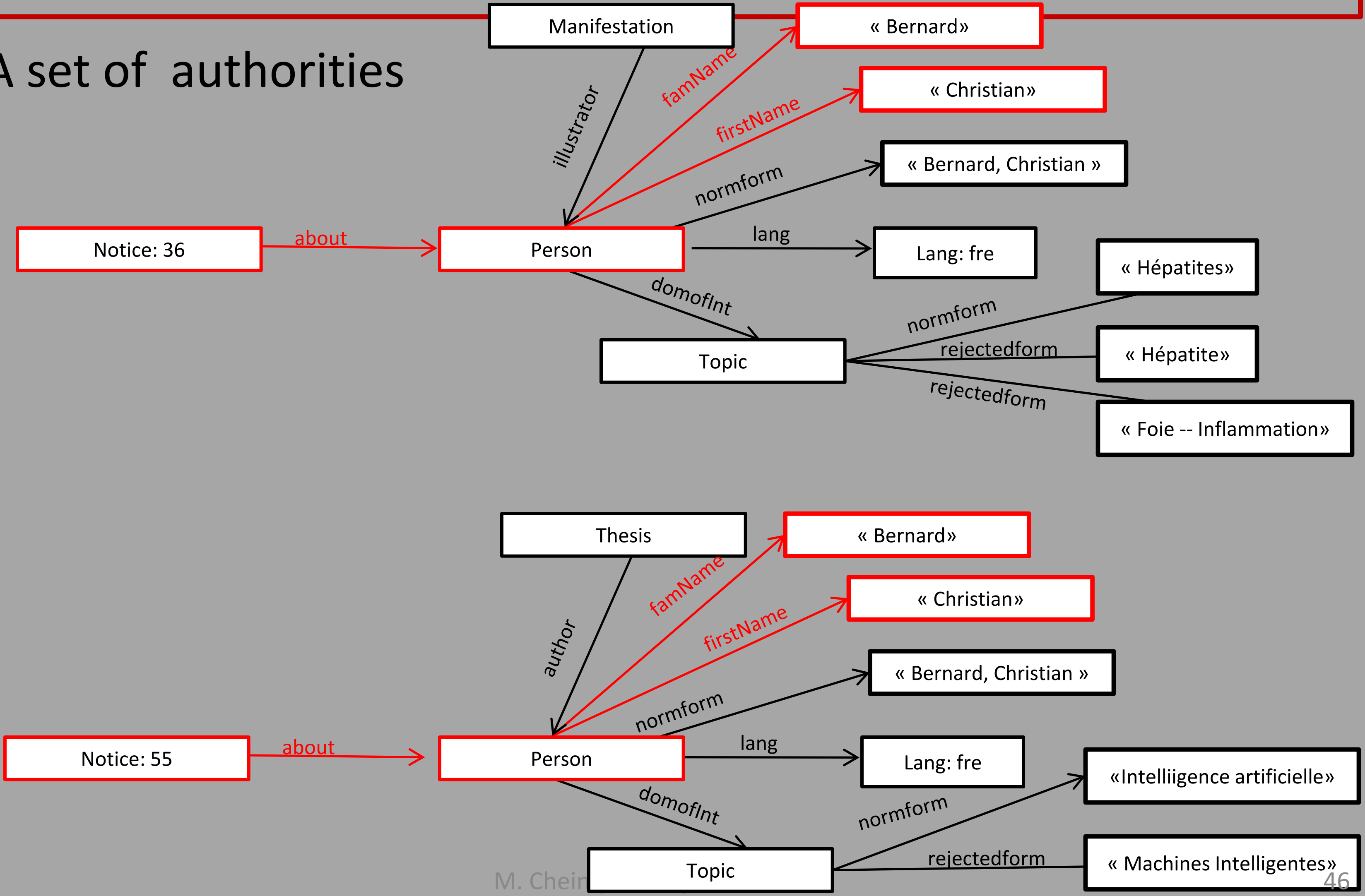


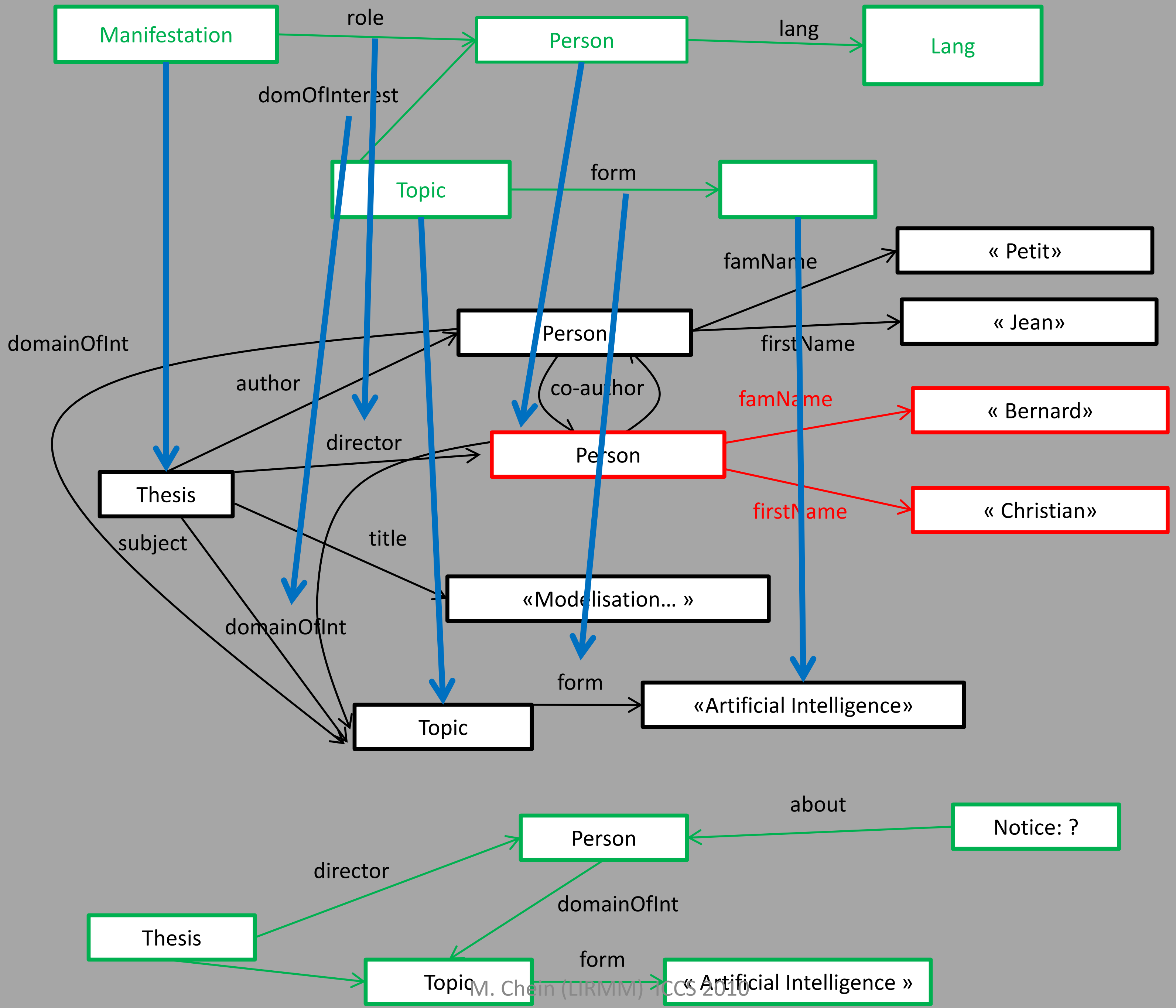
Identification Query



Identification Answers

A set of authorities



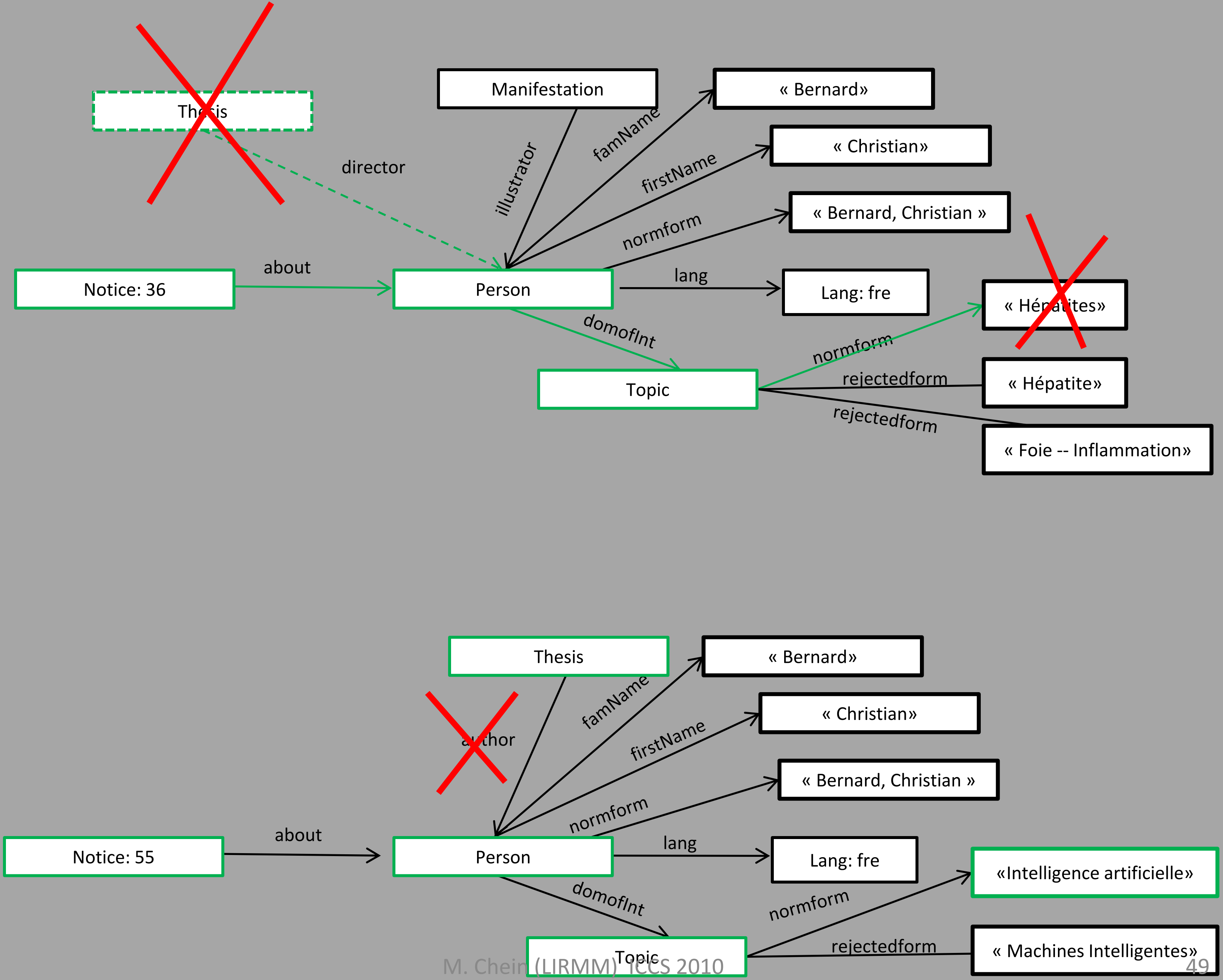


Manifestation
role
Person
lang
Lang
domOfInterest
Topic
form
famName
« Petit»
Person
« Jean»
domainOfInt
firstName
author
co-author
famName
« Bernard»
director
Person
Thesis
firstName
« Christian»
subject
title
«Modelisation… »
domainOfInt
form
«Artificial Intelligence»
Topic
about
Notice: ?
Person
director
domainOfInt
Thesis
form
Topic
« Artificial Intelligence »

Identification Ranking

The optional part is used as a ranking criteria for the selected authorities

- Complexity of the transformations of the enriched authority needed to obtain a hom from the optional part to the authority



~~Thesis~~

director

Manifestation

illustrator

famName

« Bernard»

firstName

« Christian»

normform

« Bernard, Christian »

Notice: 36

about

Person

lang

Lang: fre

domofInt

Topic

normform

~~« Hépatites»~~

rejectedform

« Hépatite»

rejectedform

« Foie -- Inflammation»

Thesis

~~author~~

famName

« Bernard»

firstName

« Christian»

normform

« Bernard, Christian »

Notice: 55

about

Person

lang

Lang: fre

domofInt

Topic

normform

«Intelligence artificielle»

rejectedform

« Machines Intelligentes»

Collective consistency checking

Constraints on relations between authorities

For an article in a journal, the publication date should be consistent with the lifedate of the author

For a thesis the domains of interest of the director and of the author should intersect

For a paper with several authors, attributes associated with co-authors of a document should respect some constraints such as lifedates, domains of interest, languages, …

Ex. A paper written by Jean Petit and Christian Bernard

«Jean Petit», ordered authorities [a11, a41, a35]

«Christian Bernard», ordered authorities [a55, a36]

If (a11,a55) and (a41,a55) and (a41,a36) are inconsistent co-authors
Return
(«Jean Petit»,«Christian Bernard»), [(a11, a36),(a35,a55),(a35,a36)]

Perspective 1 Work in progress

All notions are available in GBKR (cogui and cogitant)

- **Knowledge representation**: enrichment rules, selection pattern for each sort of authority, constraints, approximate hom and ranking
- **Experiments:** checking the methodology, the knowledge, the whole graph
- **Mapping:** other ontologies (for integrating notices coming from other sources)
- **Introduction** of the authority identification service as a tool
- **Quality of data:** de-duplication, errors in links

Perspective 2 Relationships with logic

Classical FOL

two distinct constants can have the same interpretation

KIF has two kinds of constants: for one kind, uniqueness is assumed; for the other kind, either a = b or a ≠ b is possible.

Standard names

surrogates and ordinary constants, the interpretation domain is the set of surrogates

Surrogates, ordinary constants and litterals with different boolean match predicates for different types of litterals (e.g., match_string, match_date, match_name), the interpretation domain contains the set of surrogates

Conclusion

A « killing » application?

Large amount of metadata built on standard Ontologies by professionals

The web is not a digital library …

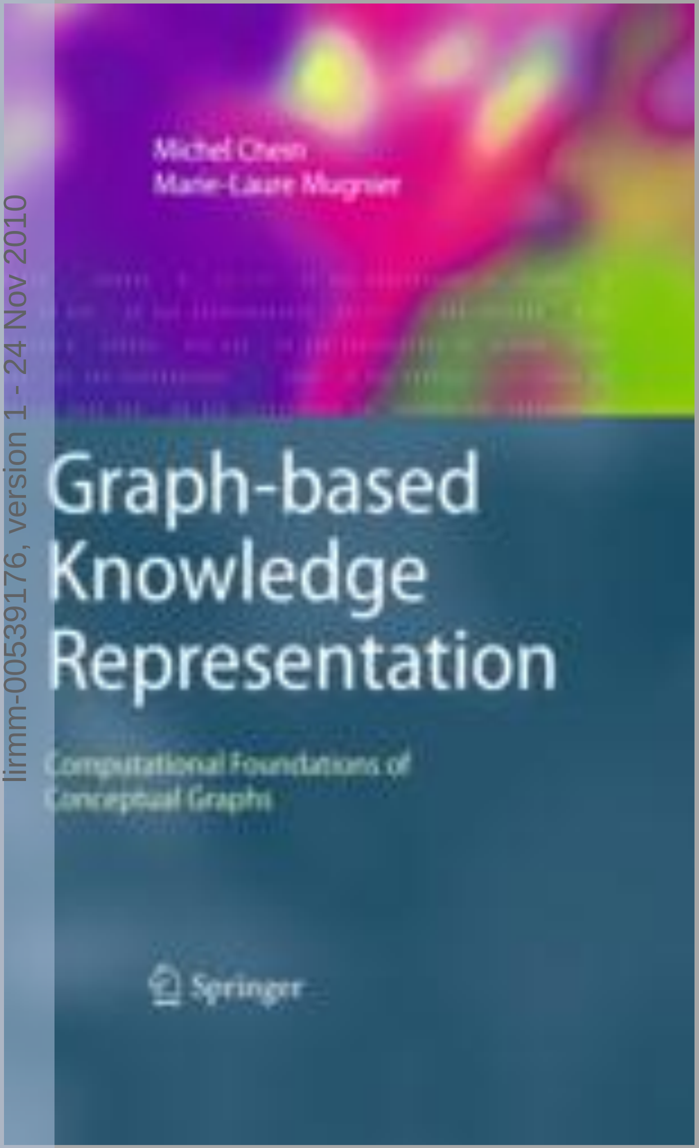
A first step before attacking part of the web?

« Pour le dire vite, le web sémantique est la revanche des métadonnées sur l'utopie du Full Text. C'est donc aussi une affaire de bibliothécaires. » Yann Nicolas

A very sensitive issue

« However, some record linkage projects (e.g., involving sensitive data on health, finance, and crime) have met with public outcries because they were perceived as secret, revealing, without-consent, inaccurate, or resulted in administrative action (U.S. General Accountability Office, 2001). Thus, it is important to keep the public in mind when embarking on an author disambiguation project. ***At the very least, author disambiguation research should be transparent***. »

Everything You Always Wanted to Know About …



Michel Chein
Marie-Laure Mugnier

Graph-based Knowledge Representation

Computational Foundations of Conceptual Graphs

Springer

http://www.lirmm.fr/gbkrbook/

Short Bibliography

O. Benjelloun, H. Garcia-Molina, D. Menestrina, Qi Su, S. Euijong Whang, J. Widom, Swoosh: a generic approach to entity resolution. The VLDB Journal, 18, 255-276, 2009

X. Dong, A. Halévy, J. Madhavan, Reference Reconciliation in Complex Information Spaces. In Proc. of SIGMOD'05, 85-96, ACM Press (2005)

D. Genest, M. Chein, A Content-search Information Retrieval Process Based on Knowledge Graphs and the Uncertainty Principle. Knowledge and Information Systems (KAIS), vol. 8, n° 3, 2005

H. Newcombe, J. Kennedy, S. Axford, and A. James. Automatic linkage of vital records. Science, 130:954-959, 1959.

F. Saïs, N. Pernelle, M.-C. Rousset, L2R: a Logical Method f *In proceedings of the Twenty-second AAAI Conference on Artificial Intelligence (AAAI-07) pages 329-334, Vancouver, British Columbia, Canada.* or Reference Reconciliation

N. R. Smalheiser and V. I. Torvik, Author Name Disambiguation, in Volume 43 (2009) of the Annual Review of Information Science and Technology (ARIST) (B. Cronin, Ed.)

K. Smith-Yoshimura. Networking Names. Report produced by OCLC Research. 2009

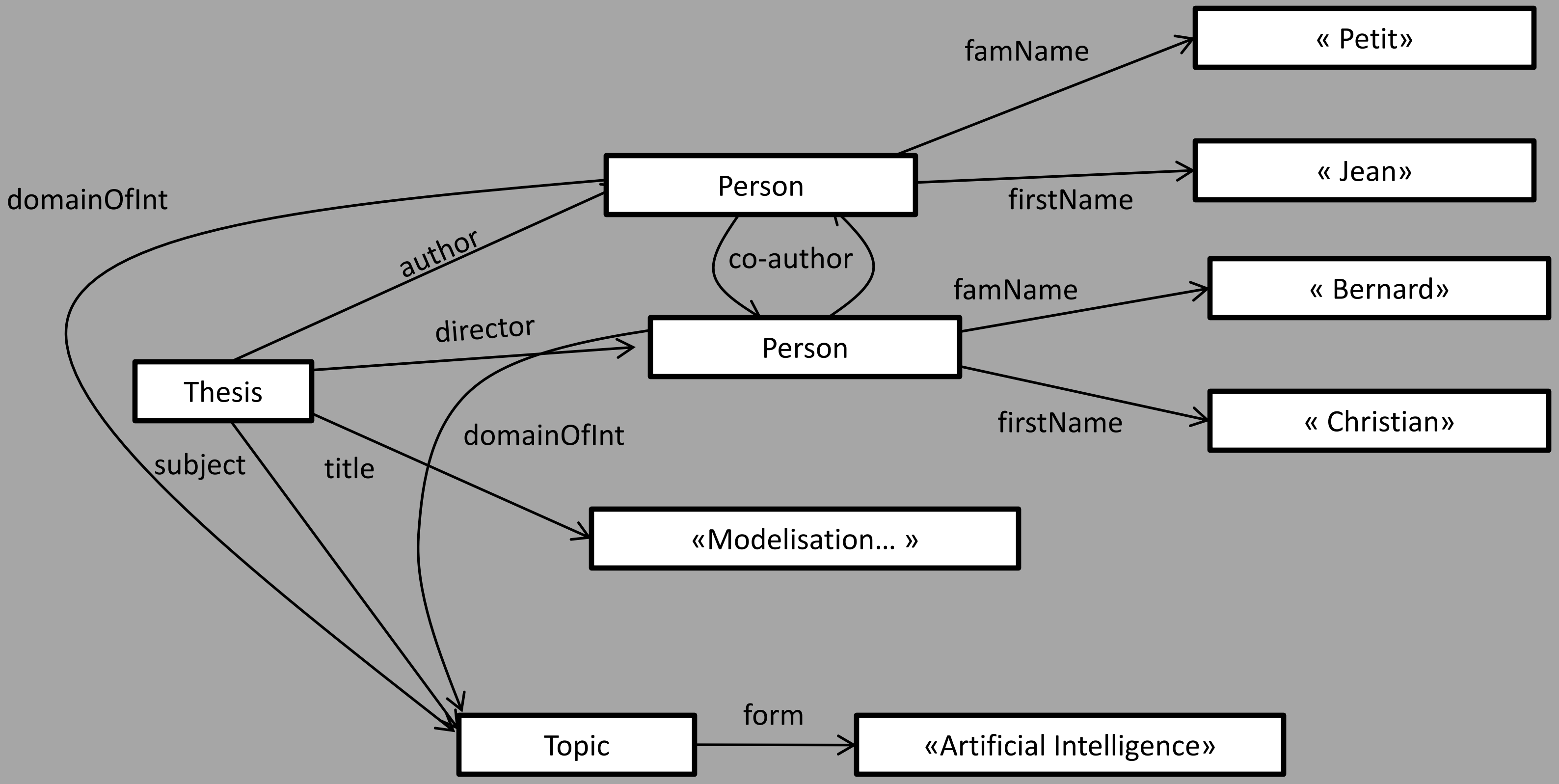
Published online at: **http://www.oclc.org/programs/reports/2009-05.pdf.**

John F. Sowa, *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks Cole Publishing Co., Pacific Grove, CA, ©2000. Actual publication date, 16 August 1999.

Thank you for your attention!

Identification Selection

Choose the selection pattern



What’s In a Name?

'What is in a name? Very much if the wit of man could find it out.' Whoever penned this well known saying undoubtedly had it each with a history behind it.