From publications to knowledge graphs

Panos Constantopoulos, Vayianos Pertsas

Athens University of Economics and Business, Department of Informatics

The 13th International Workshop on Information Search, Integration, and Personalization,

9-10 May, 2019, Heraklion

The wider aim

A process-oriented approach to supporting research data sharing and open science

Consider the questions:

Has a particular research question been addressed and how?

Who has worked on a particular topic and what is known about their work?

Which projects has a given method been used in?

Which are the preferred tools for a certain kind of work?

How has a particular experiment that uses a specific method been conducted?

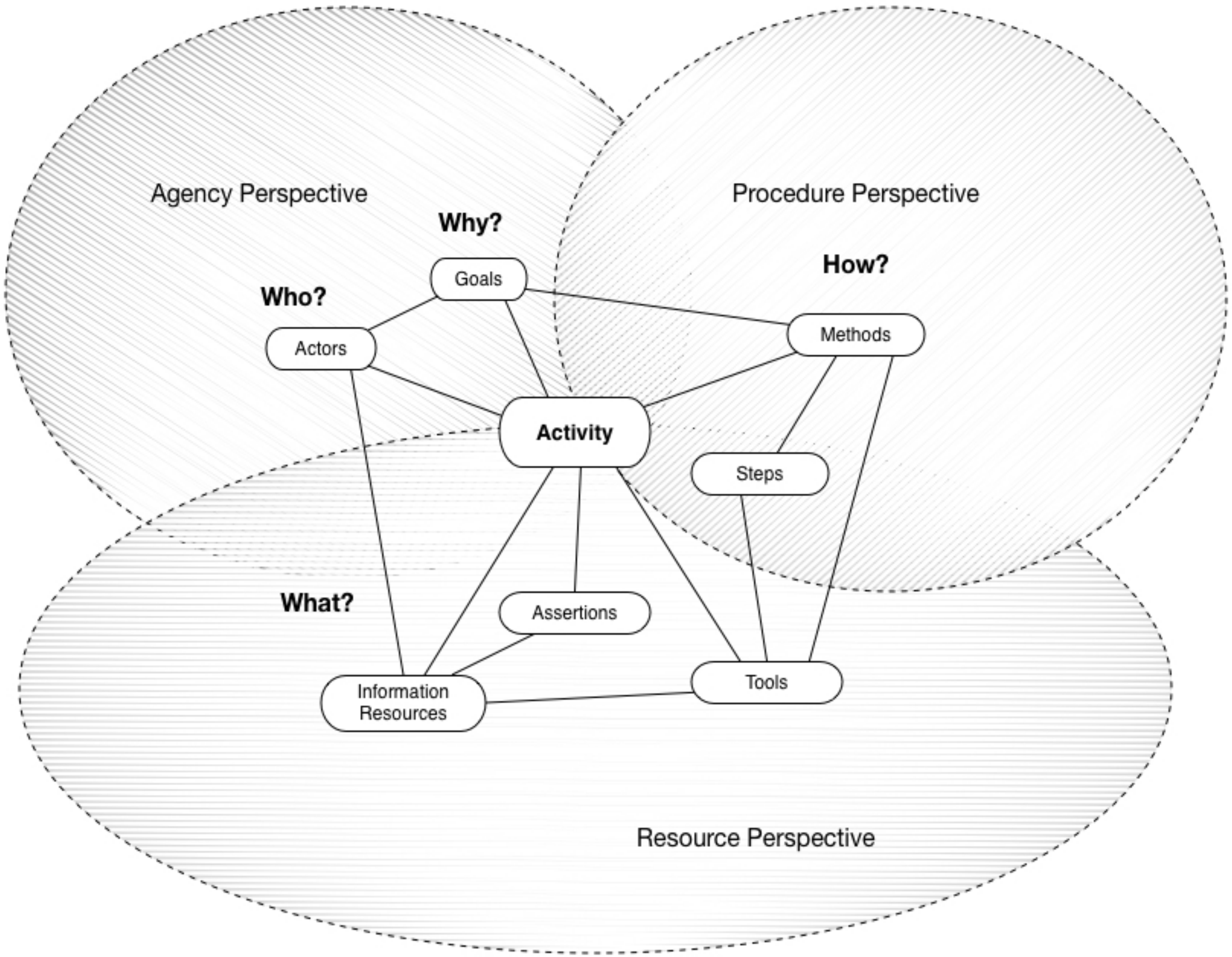
Answering these questions today:

1. Use search engines, consult specialized sources
2. Find relevant publications
3. Read them and find out about
   - research activities described
   - methodology followed
   - goals, questions, topics addressed
   - results produced
   - resources and tools used, etc.
4. Find and use other relevant resources (e.g. images, tools, repositories, etc.)
5. Combine all of the above, and continue

The Scholarly Ontology (SO) :

Captures knowledge about scholarly work so that we can answer questions of the form:

“Who does what, where, when, why and how…?



The Scholarly Ontology (SO):

Framework for documenting research practice.

Supports leveraging Linked Data.

Provides a layered architecture that supports interconnection /compatibility with foundational ontologies.

Admits domain-specific extensions.

Models research processes through different perspectives, covering the entire spectrum of scholarly work.

Extends the -domain specific- NeDiMAH Methods Ontology (NeMO).

NeDiMAH: Network for Digital Methods in the Arts and Humanities, ESF Research Network, 2011-15

- Researching digital methods in arts & humanities
- A collaborative forum of communities of practice

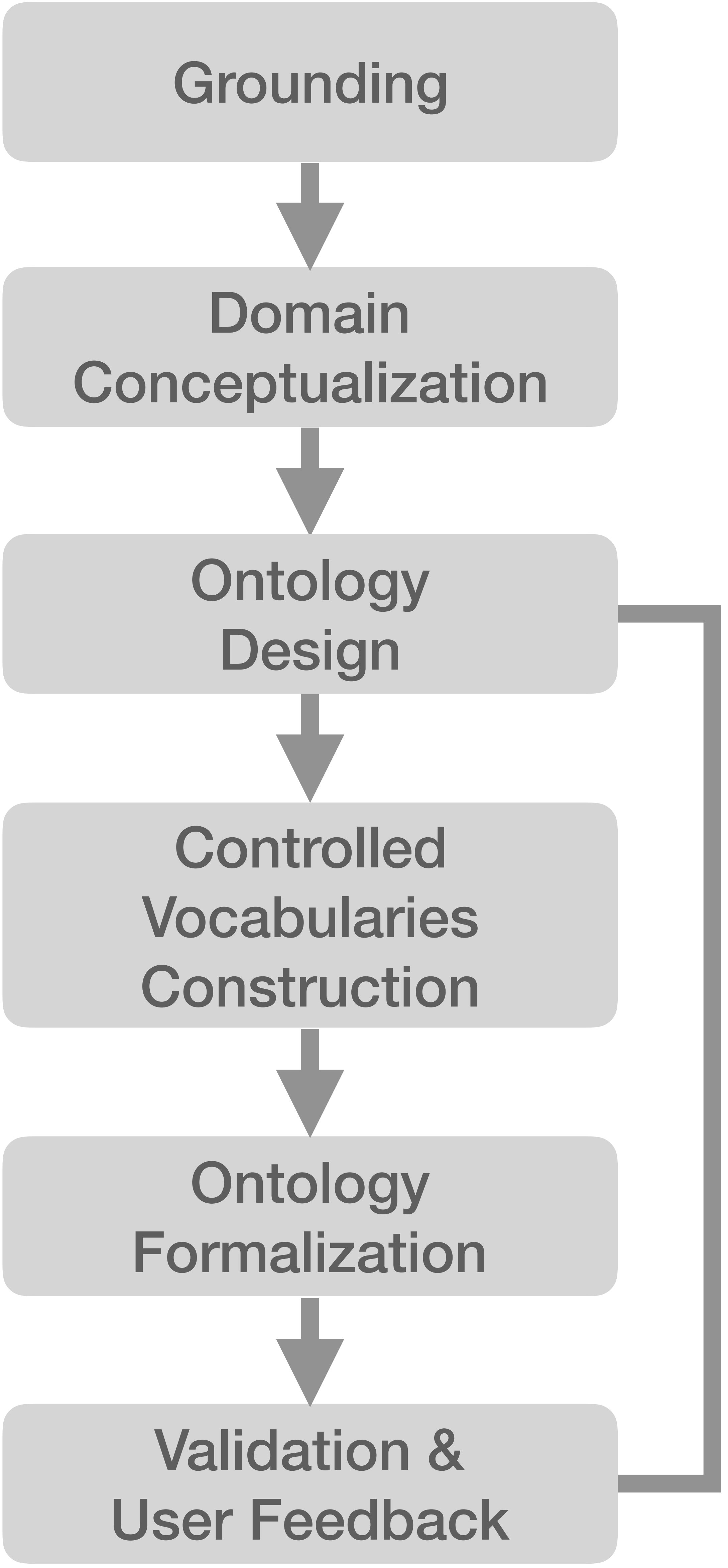
www.nedimah.eu

The NeDiMAH Methods Ontology - NeMO:

- A formal ontology for Digital Humanities, including classification and a shared vocabulary
- Incorporates existing DH taxonomies (e.g. Oxford ICT, TaDiRAH, DHCommons)
- CIDOC CRM compatible
- Contributed to ESF Report: Research Infrastructures in the Arts and Humanities

Why use an ontology?

- Provides a formalization of basic concepts.
- Provides a conceptual framework for complex query answering.
- Acts as semantic glue between different taxonomies.
- Supports the development of an ecosystem of interoperable resources and services for discovering, understanding, selecting, linking and contributing content, tools and methods.



Ontology Development:

Empirical research using semi-structured interviews with scholars from across Europe (earlier work).

Leverage related work: AHDS computational methods taxonomy, TaDiRAH, Scholarly Research Activity Model (Preparing DARIAH, EHRI), ARIADNE, Europeana Cloud, SPAR/CiTO, EXPO/CRM-Sci, etc.

Analysis of the ground evidence, core concepts and relationships of the domain identified. Modelling decisions.

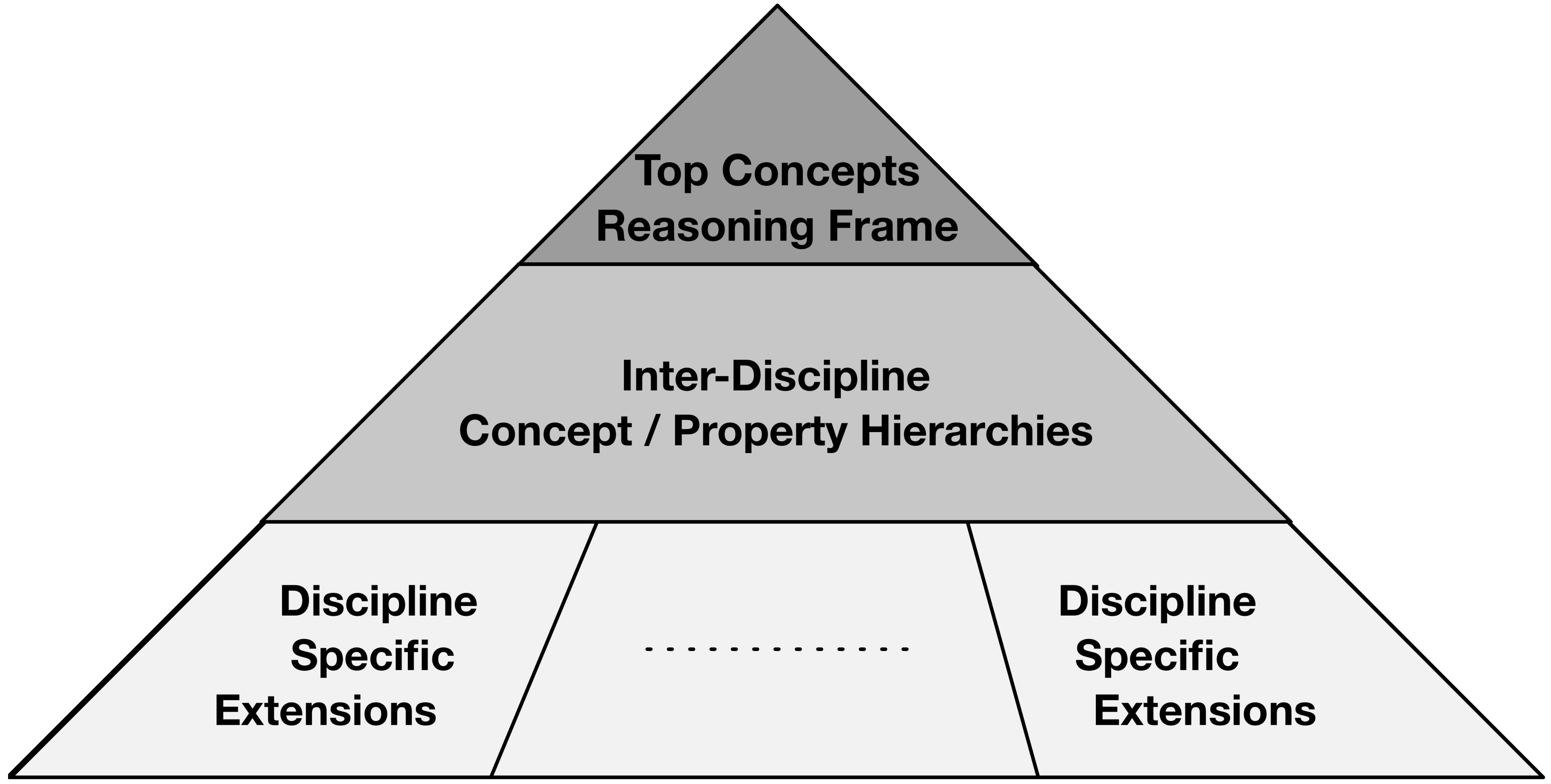
Definitions in textual form, examples and mappings of SO terms to and from terms of other taxonomies.

Encoding in RDFS and SKOS (where needed).

Workshops: validation, collection of use cases and information needs.

Scholarly Ontology (SO):

a 3-layer structure

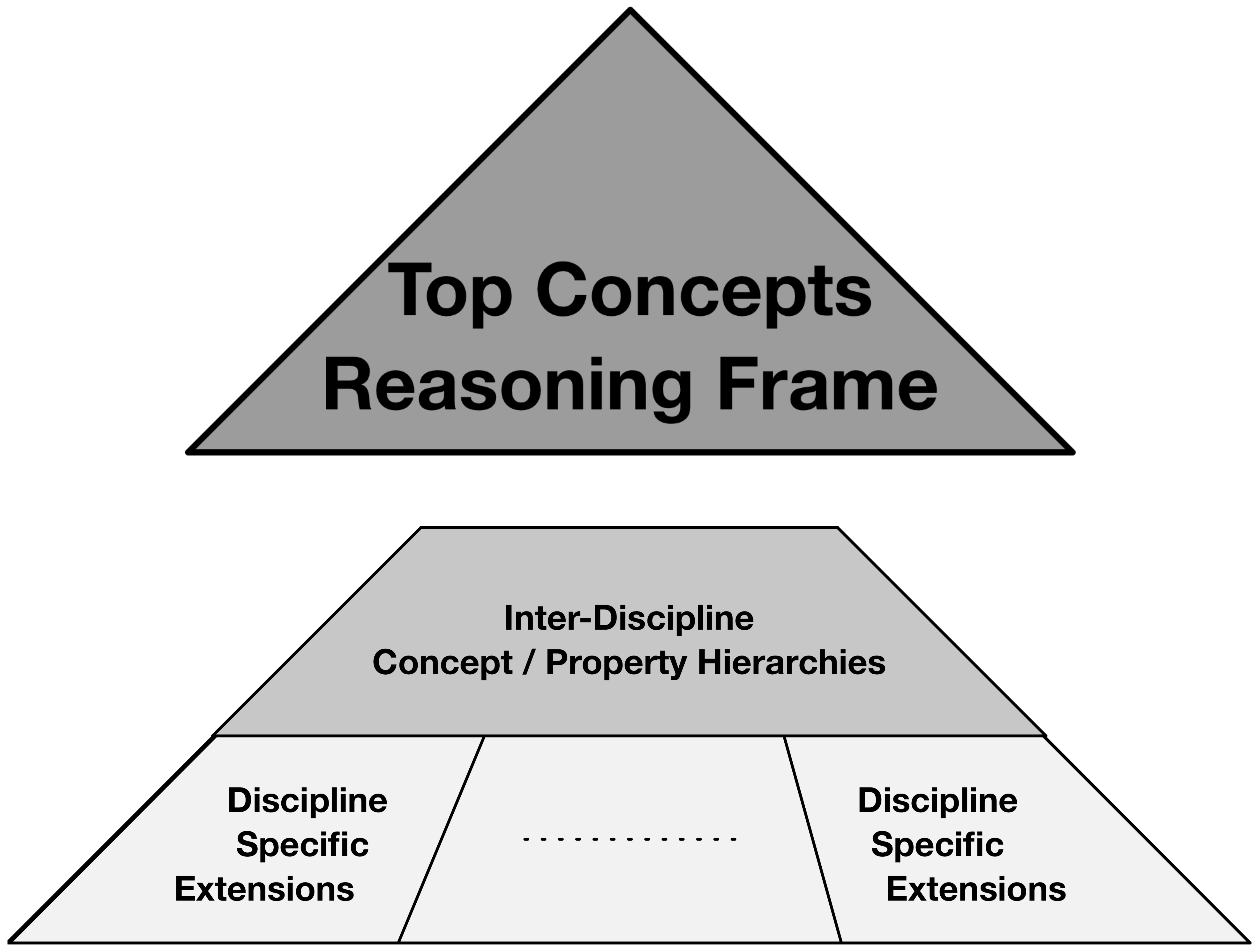


Top layer:

most general concepts and properties

frame of reference

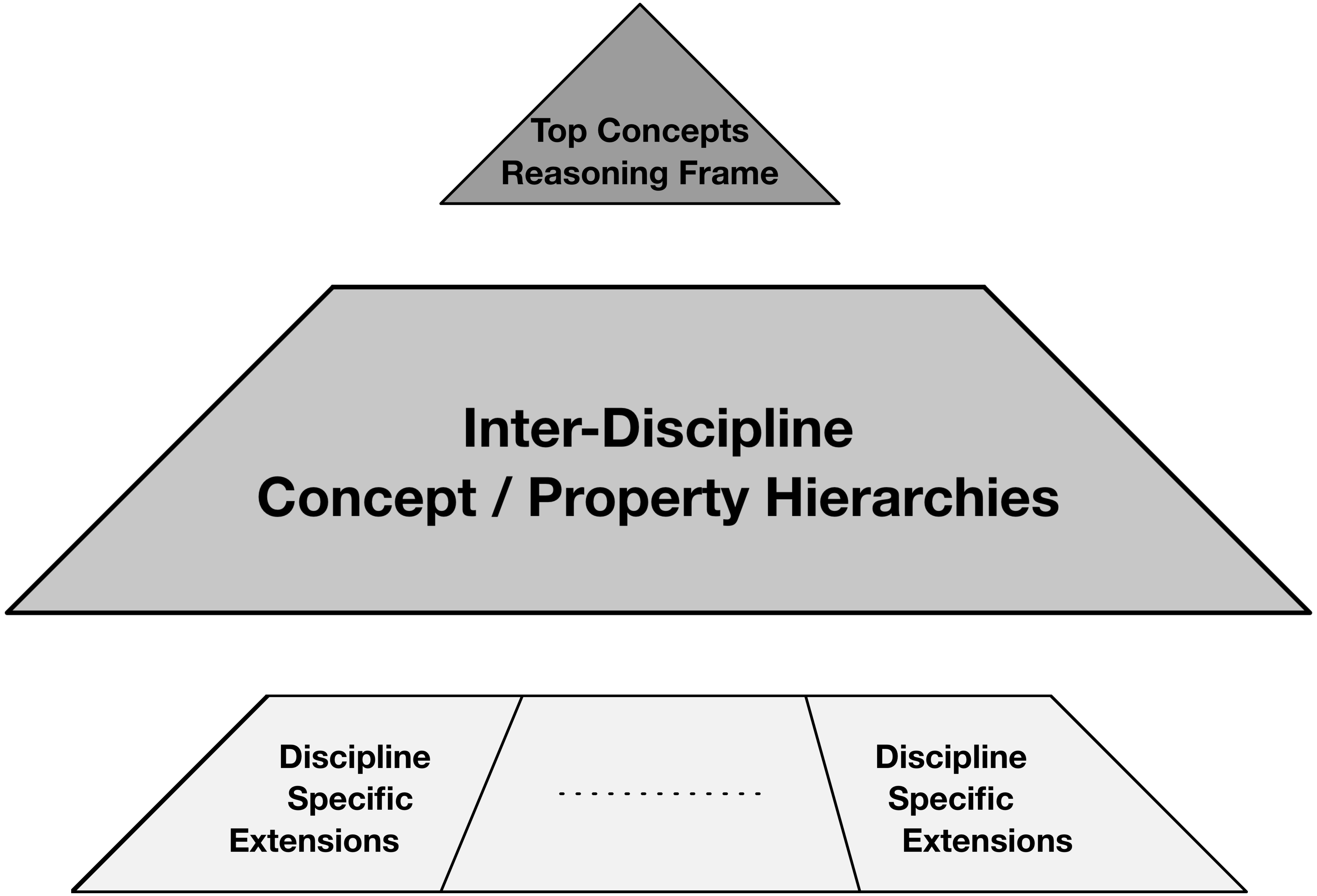
basis for linking with reference ontologies



Middle layer:

generic aspects of research processes

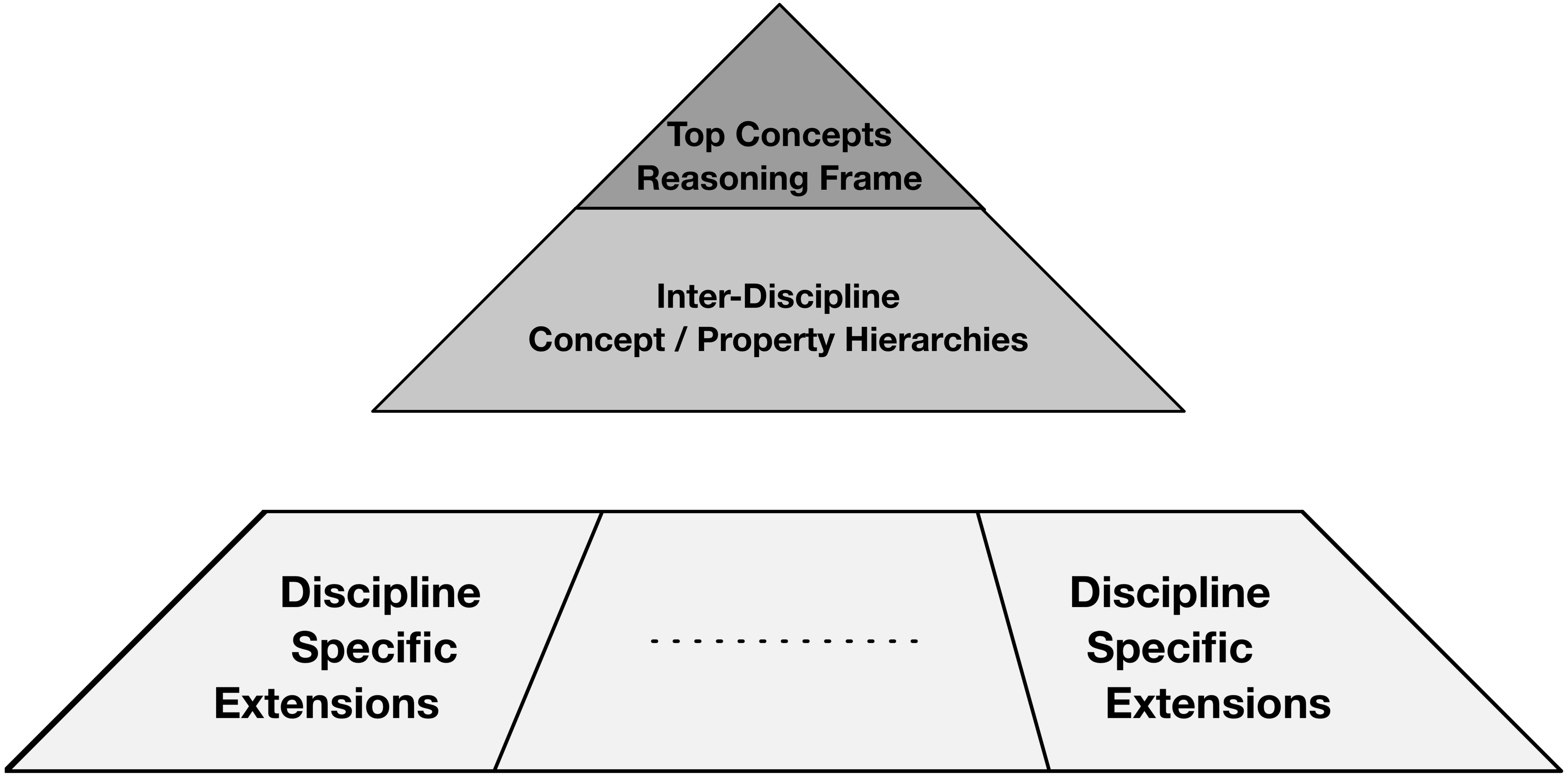
common across disciplines



Bottom layer:

fine-grain aspects of research practices

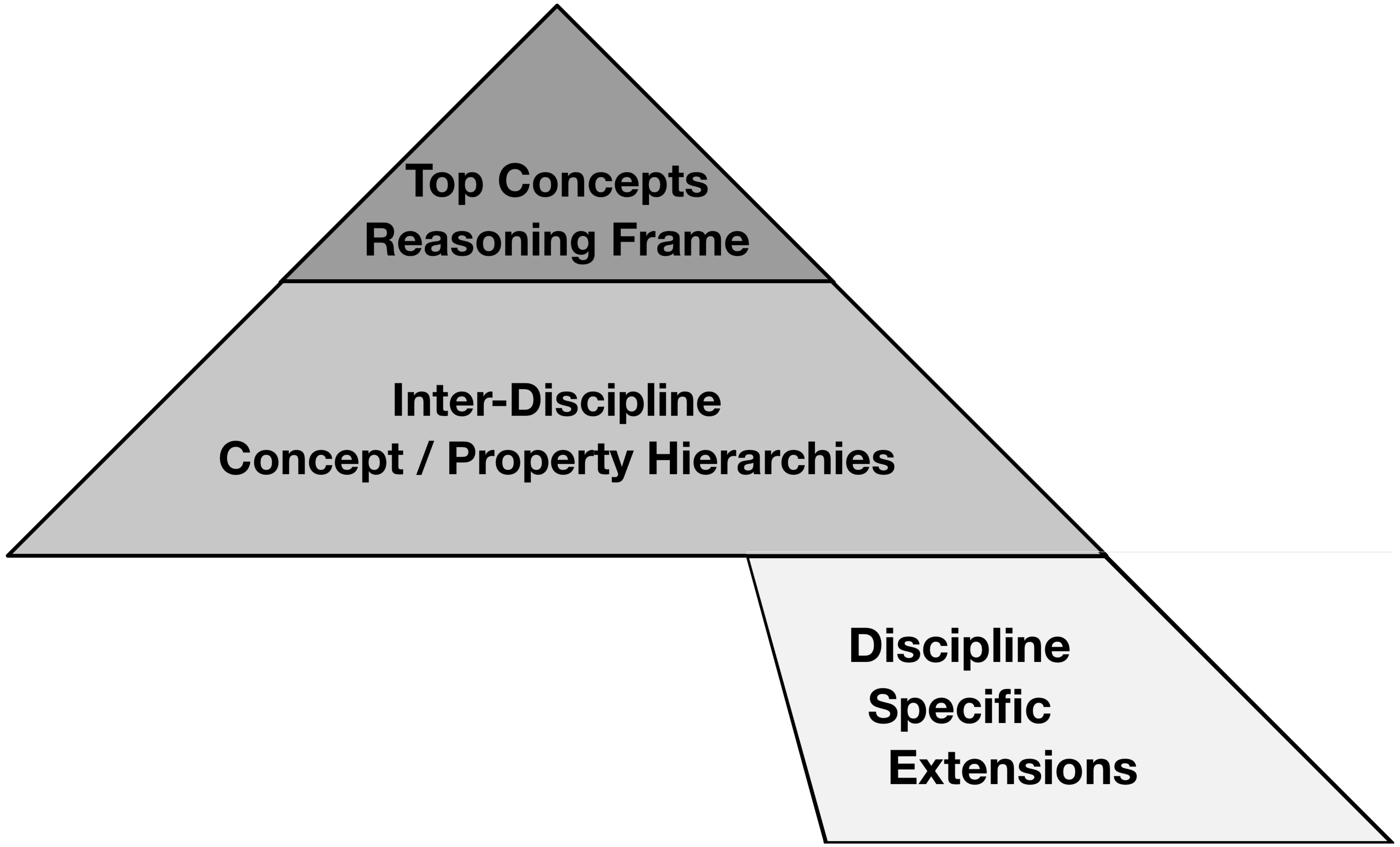
discipline-specific



A domain ontology

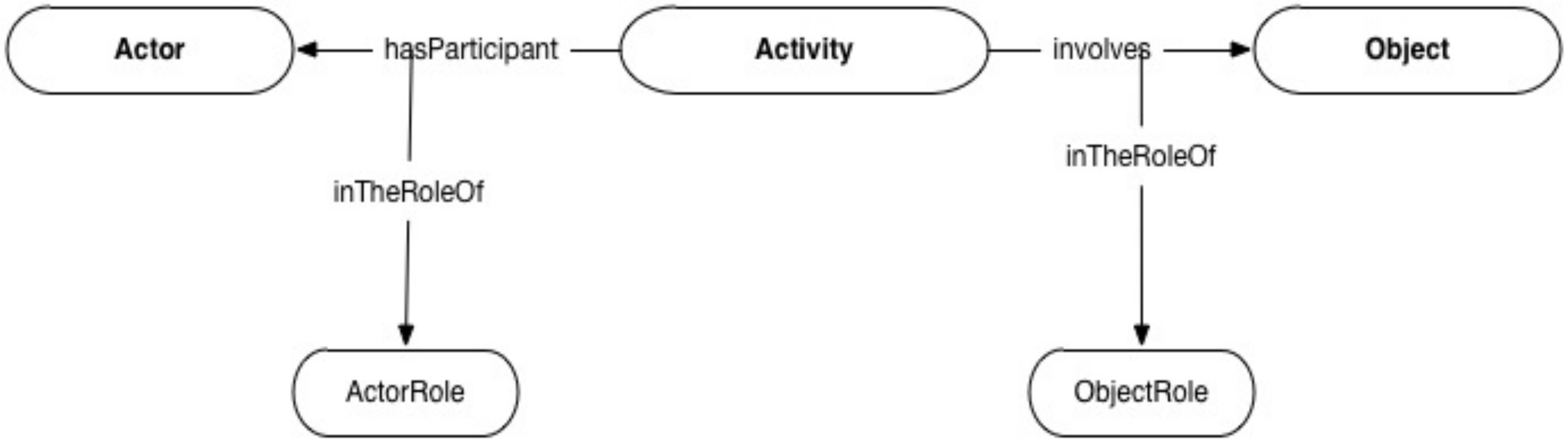
=

top layer
+ middle layer
+ discipline specific extension



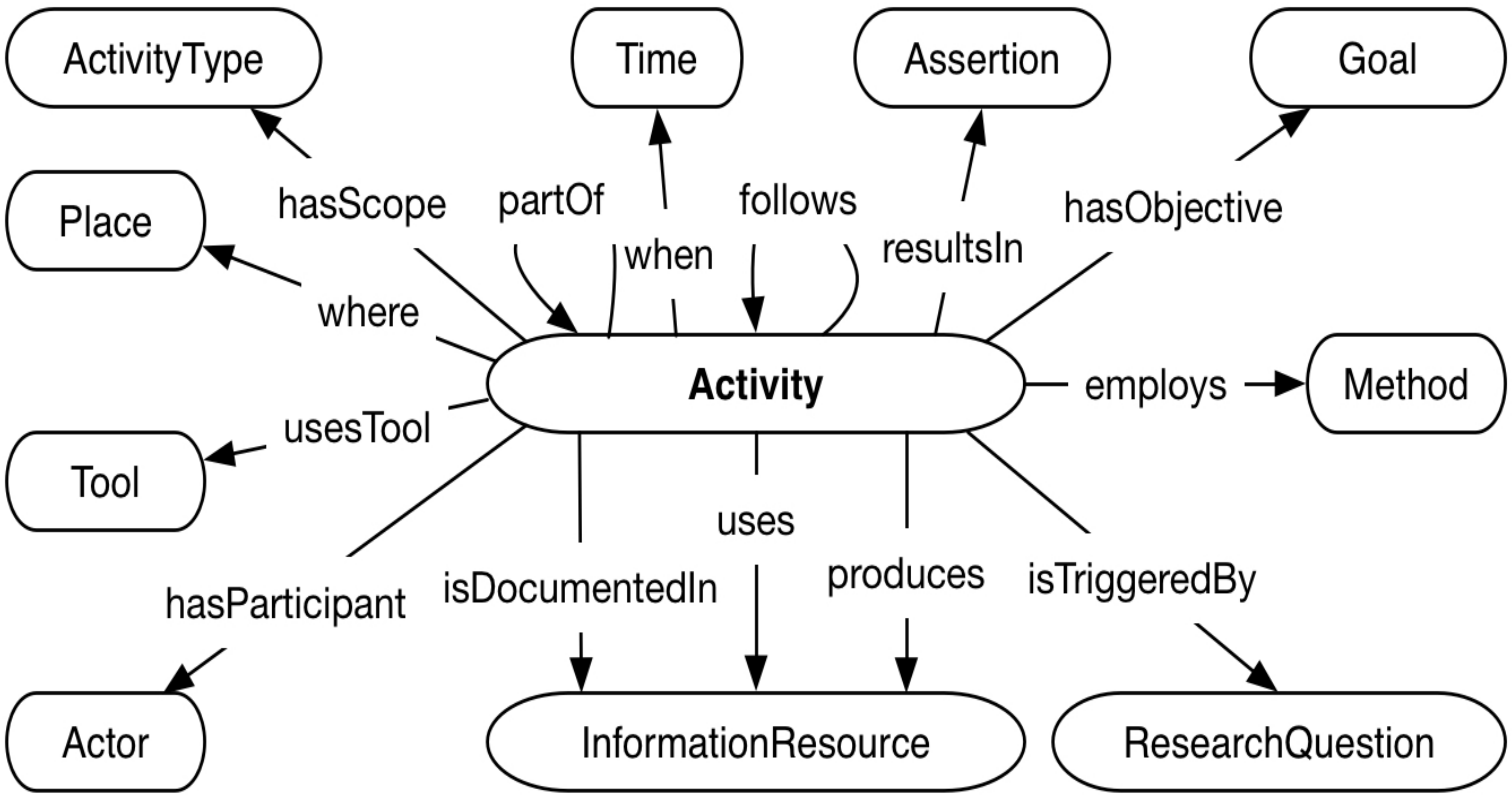
SO:

Top Elements



Activity Perspective:

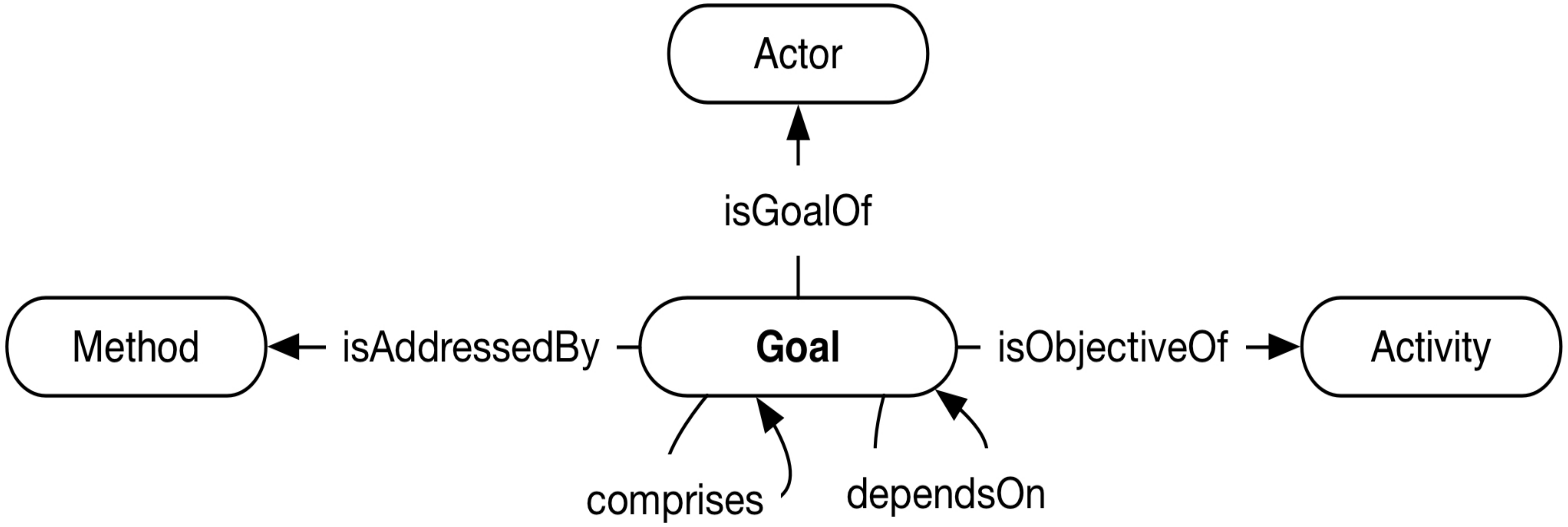
Activity:
Deliberate acts that have been carried out
(e.g. experiments, excavations, evaluations etc.)



Agency Perspective:

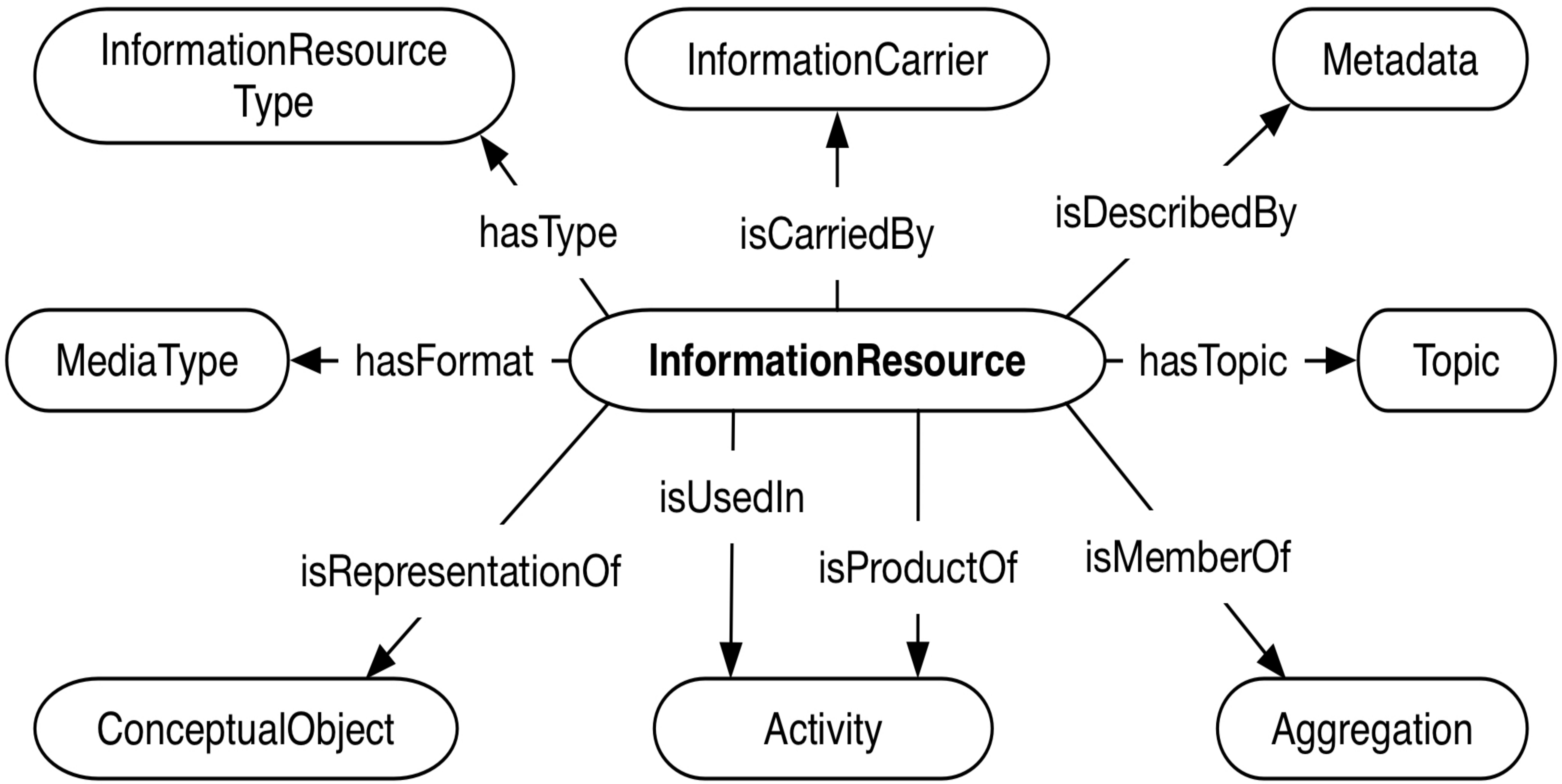
Goal:
Assertions representing explicit research goals

Actor:
Persons
Groups / Organizations



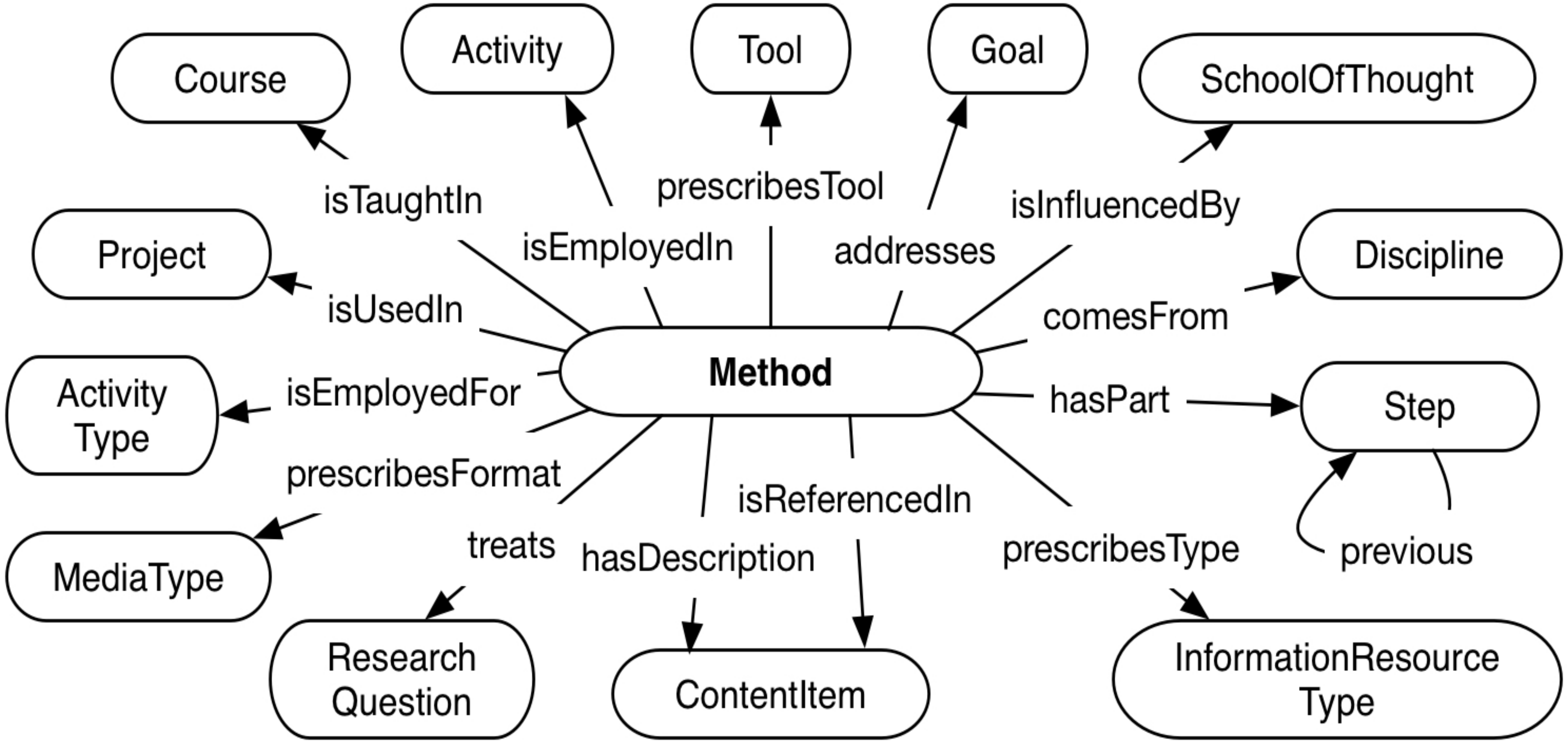
Resource Perspective:

Information Resource: Concrete manifestations of conceptual objects (e.g. research article, map, image, dataset,…)



Procedure Perspective:

Method:
Prescribes how to perform a specific act (Activity) (e.g. HDR photography, macro photography)



Grounding and Evaluation:

Based on about 100 questions gathered from different researchers.

Examples:

"Given a specific goal, retrieve all research activities that deal with it using machine learning methods."
E.g. Perform stylistic analysis

"List the tools used in more than one activity employing methods which concern a particular research topic and come from either Computer Science or Linguistics."
E.g. Computational Stylistic Analysis

Procedure:

- Each question was analyzed into the given and requested facts and transformed to a SPARQL query.
- Each fact was mapped to corresponding classes / relations of SO.
- Evaluation was based on the % coverage of query concepts by ontology concepts.

Results:

- 97% coverage of the questions (errors were mainly due to unclearly formulated questions)
- 82% of the questions correspond to direct link queries

Activity

Activity type

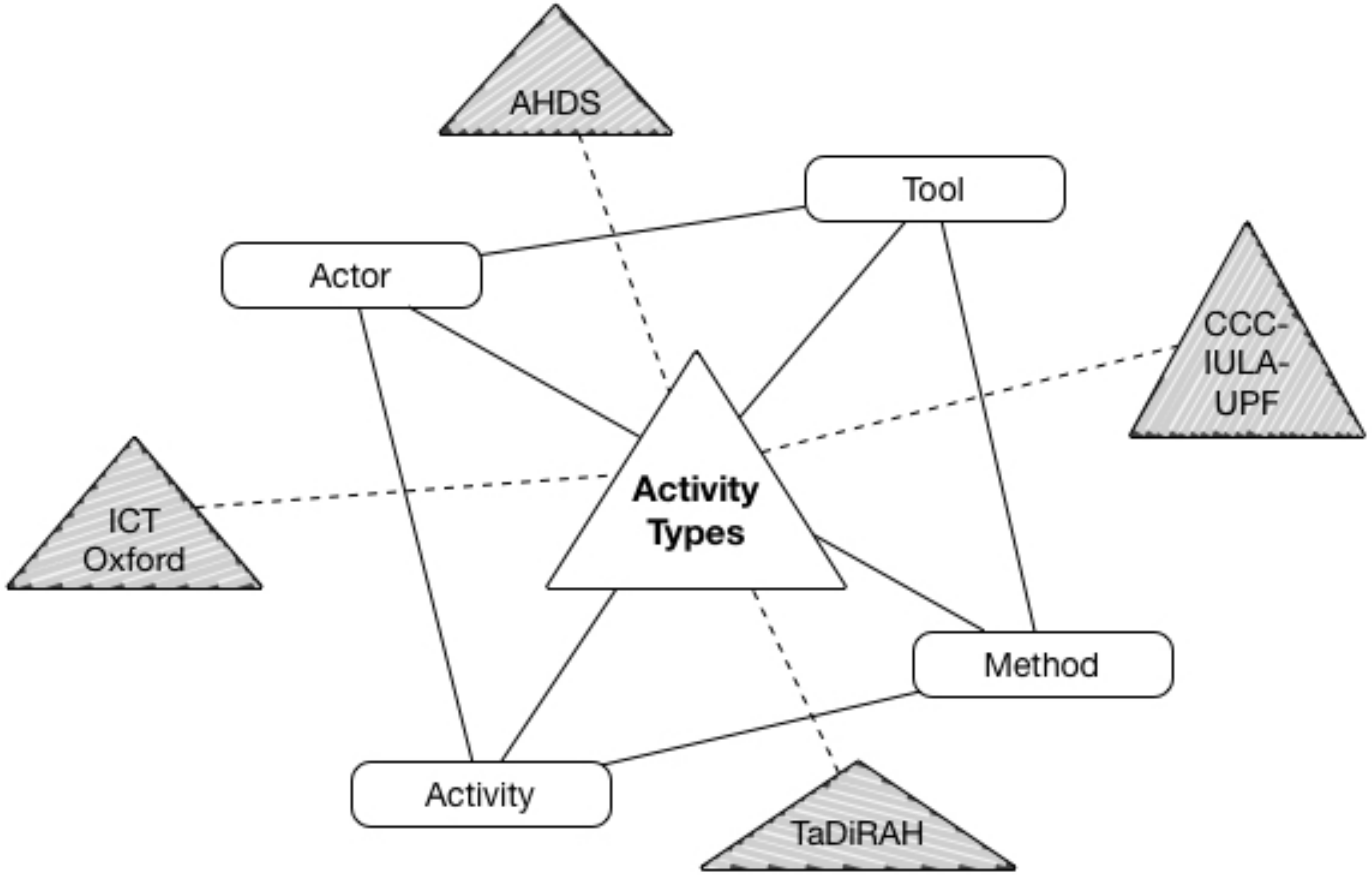
Method

Activity	Activity type	Method
Translating Herodotus Histories	Translating	Semantic translation Idiomatic translation
Creating a digital collection of letters from WW1	Collecting	Crowdsourcing
Creating an annotated corpus of poems from WW1	Annotating	POS tagging
Photographing Louvre sculptures	Photographing	HDR photographing Macro photography

Activity types:

- Denote the nature of activities
- Organized as a taxonomy
- Provide semantic context for relations
- Serve as index terms for retrieval
- Function as a “gateway” through which other taxonomic structures can be imported/mapped

.

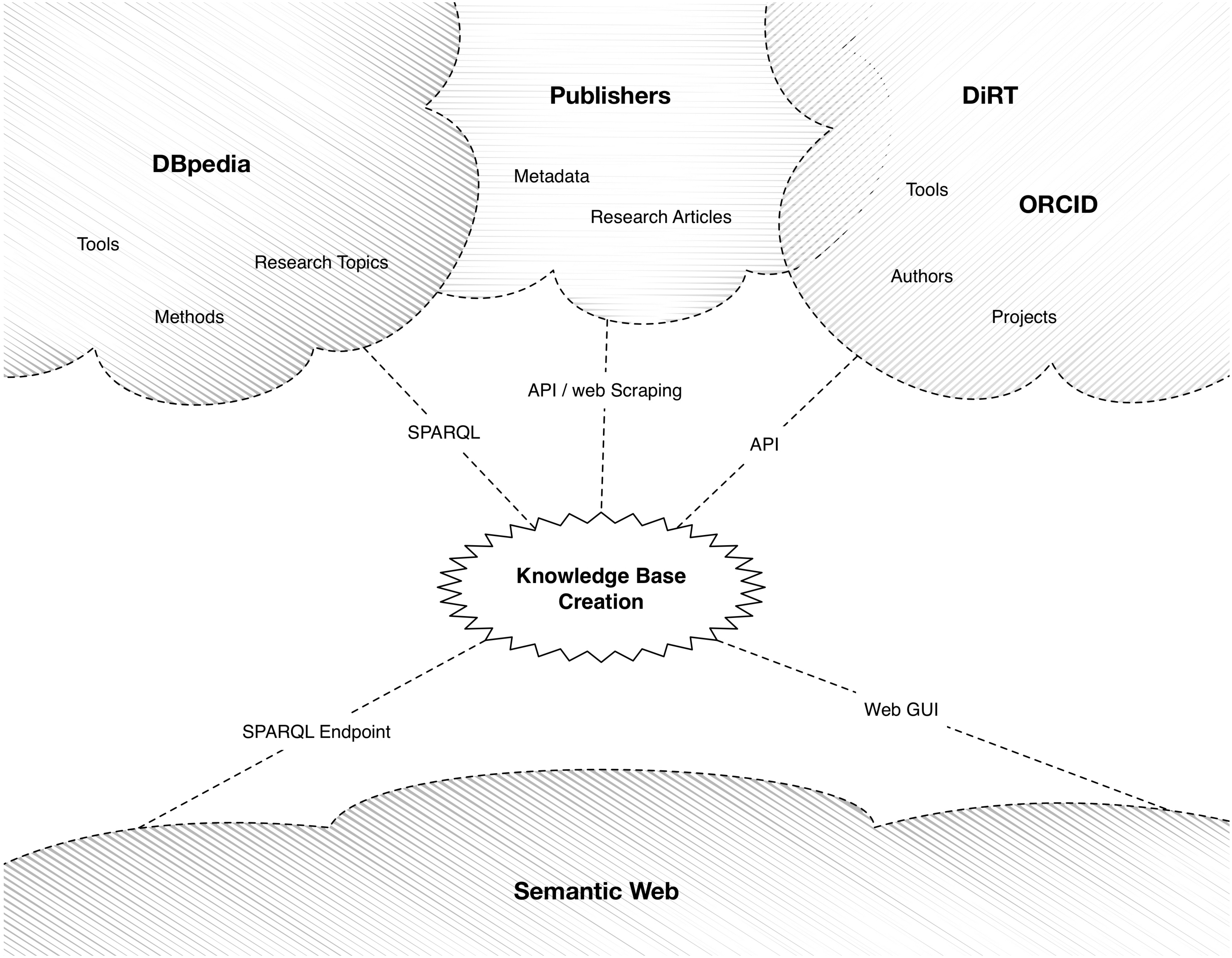


Knowledge base creation

Research Spotlight

Approach:

- Harvest repositories and websites
- Extract metadata
- Extract information from text
- Populate SO Classes
- Publish as linked data



Information extraction from publications

Challenges:

Information from publication metadata needs to be exploited.

Named entities of non-common type (such as research methods) need to be recognized from plain text.

Non-named entities (such as activities, goals, propositions) need to be identified and extracted from plain text.

Extracted entities need to be interrelated according to their semantics.

All extracted information needs to be aligned in a semantic framework for comparison or integration with other existing knowledge published as linked data .

Knowledge base creation

Stage 1

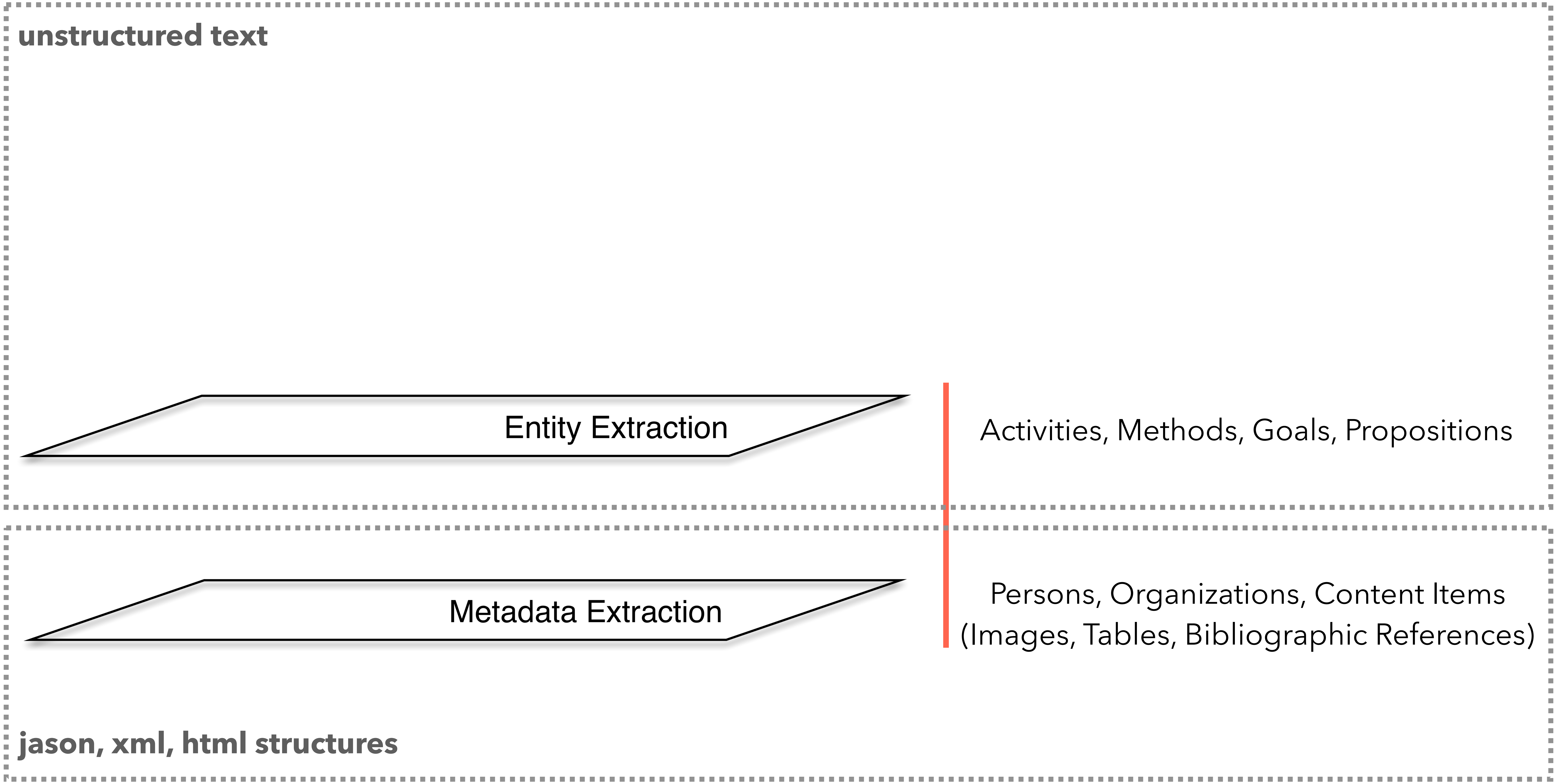
Metadata Extraction

Persons, Organizations, Content Items
(Images, Tables, Bibliographic References)

jason, xml, html structures

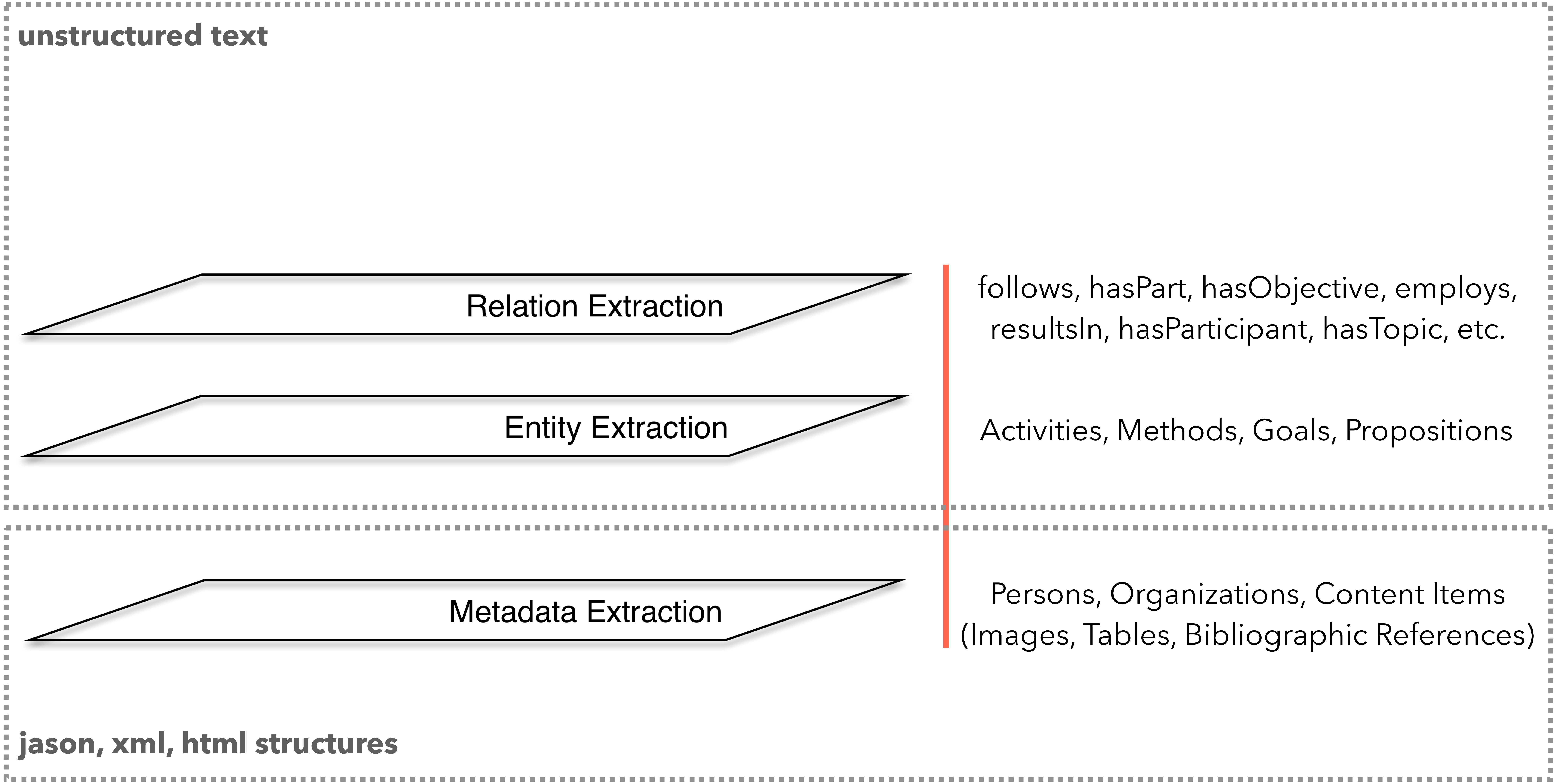
Knowledge base creation

Stage 2



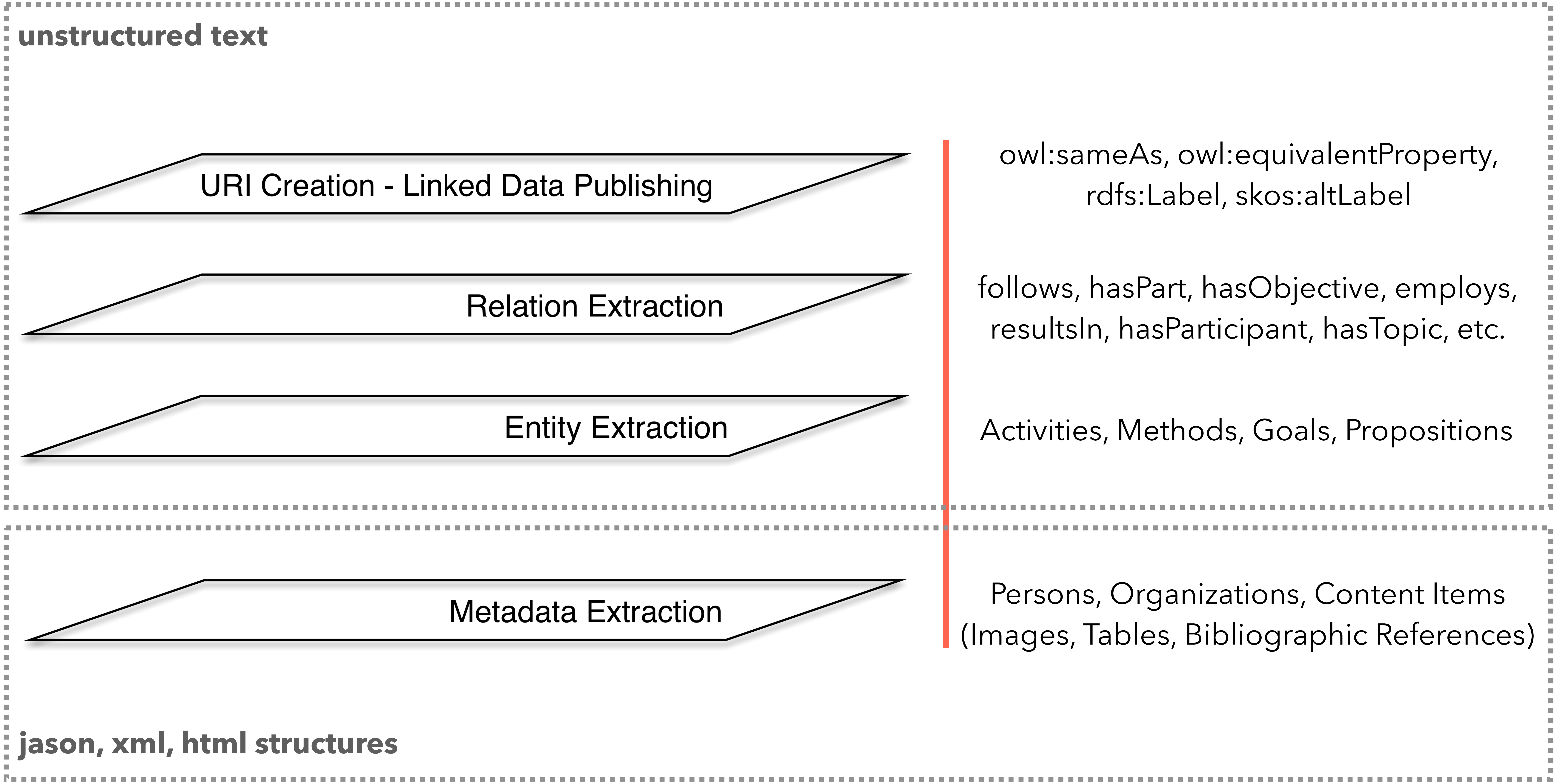
Knowledge base creation

Stage 3



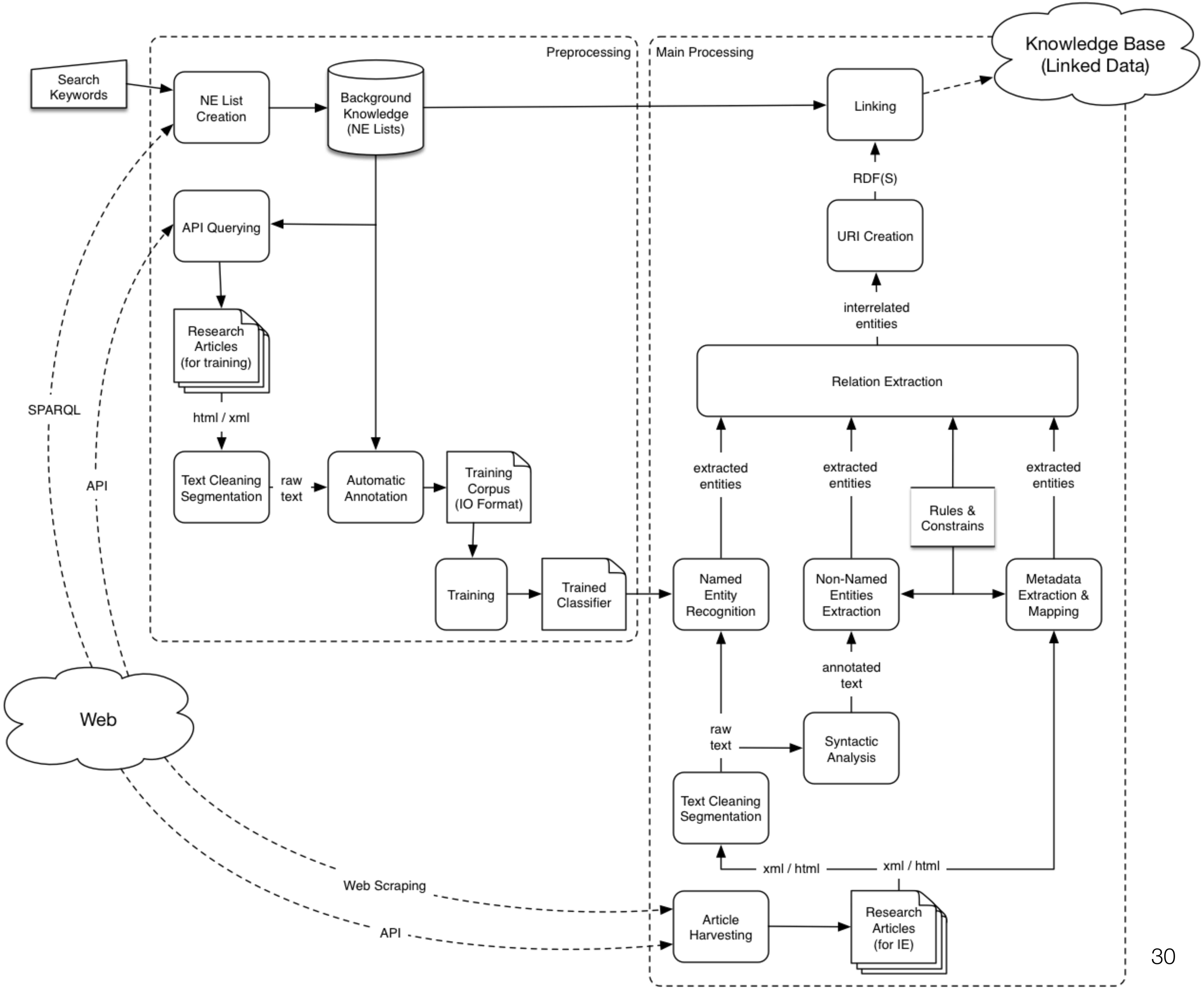
Knowledge base creation

Stage 4



Knowledge base creation

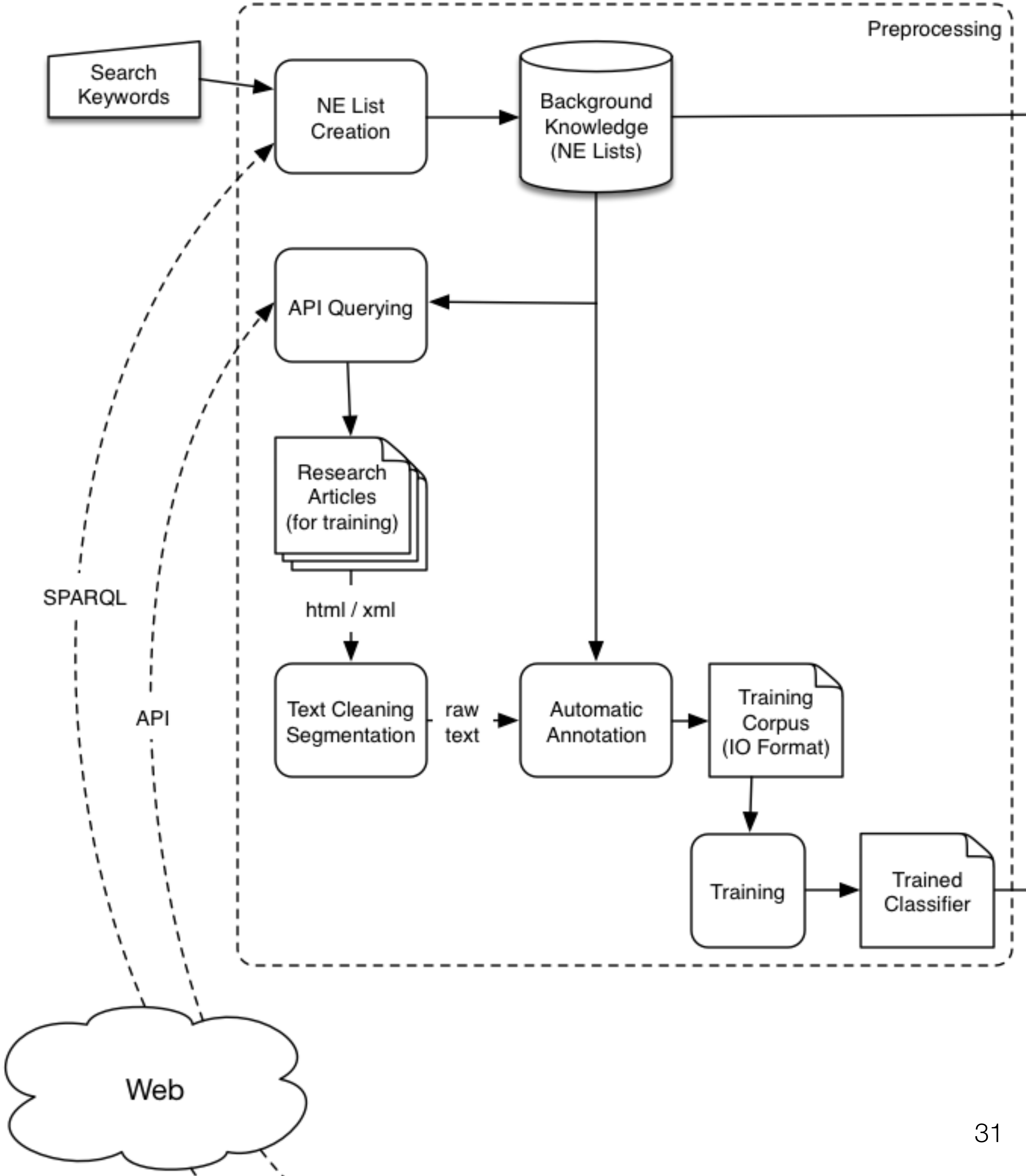
The process:



Knowledge base creation

Preprocessing:

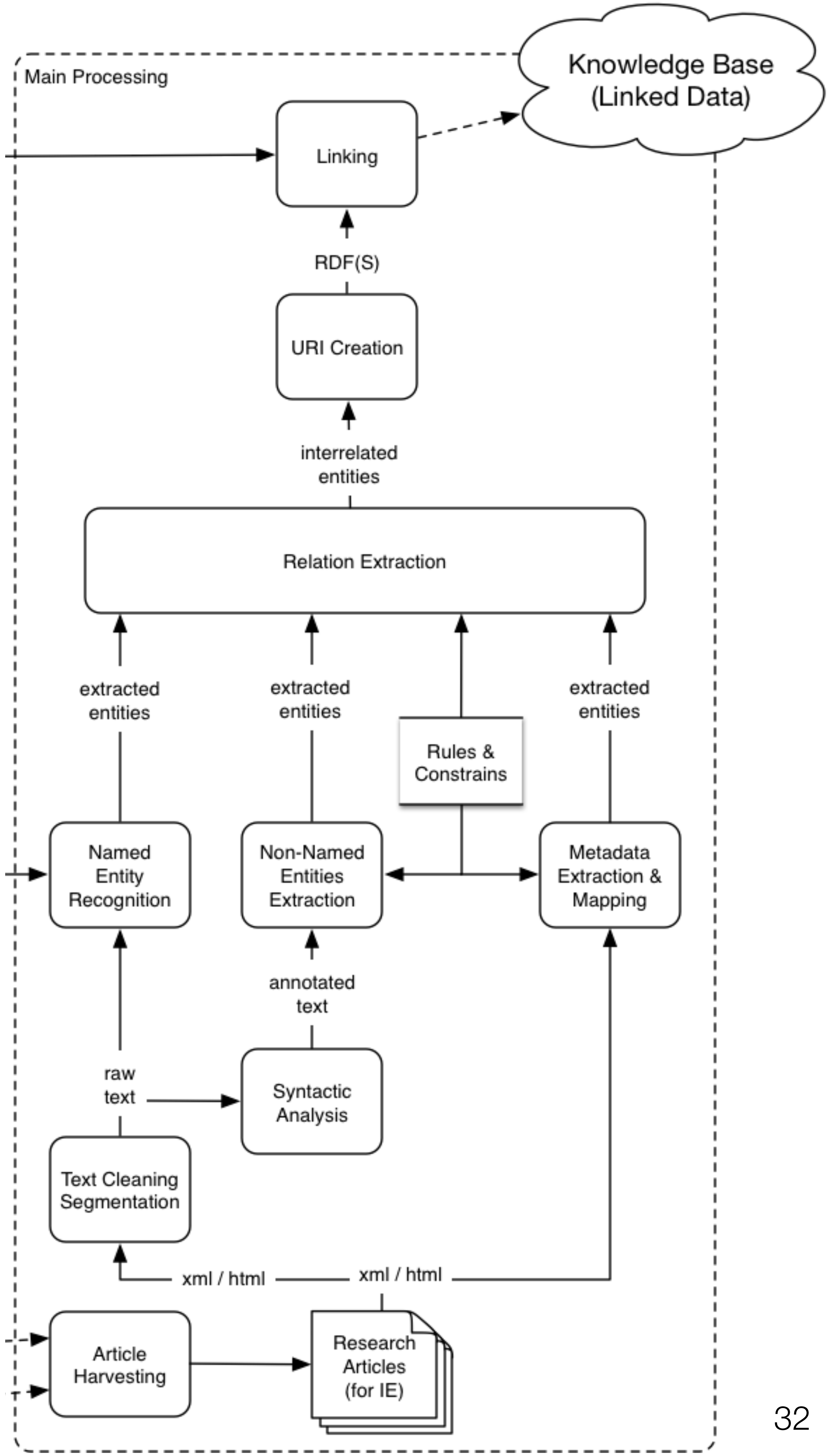
- Use DBpedia for creating lists of NE (Methods, Topics)
- Harvest research articles and use the NE lists for distant supervision



Knowledge base creation

Main processing:

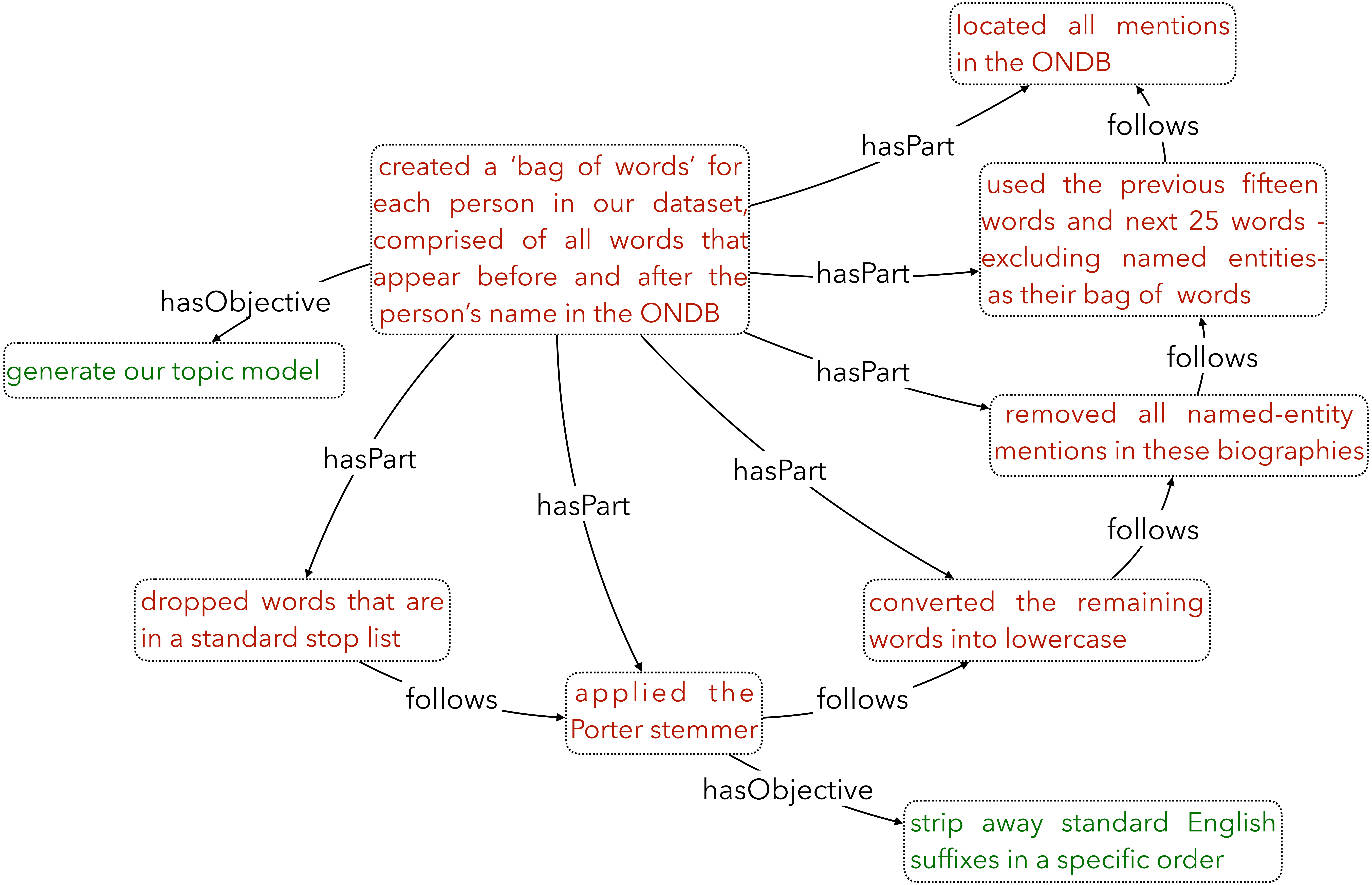
- Harvest research articles for IE
- Extract Metadata
- Extract Named Entities
- Extract Non-Named Entities
- Extract Relations
- Create URIs
- Link with other Linked Data
- Access to Knowledge Base through Web Interface / SPARQL Endpoint



Example:

To generate our topic model, we created a ‘bag of words’ for each person in our dataset, comprised of all words that appear before and after the person’s name in the ONDB. Specifically, for each person in the network, we located all mentions in the ONDB, and used the previous fifteen words and next 25 words -excluding named entities- as their bag of words. We then removed all named-entity mentions in these biographies and converted the remaining words into lowercase. Next we applied the Porter stemmer, in order to strip away standard English suffixes in a specific order. For example, the Porter stemmer turns the word ‘publisher’ into ‘publish’, and does same to the word ‘published’. We then dropped words that are in a standard stop list - which includes words like ‘and’, ‘the’, etc. -provided in the text-mining R package tm.

Example:


located all mentions in the ONDB
follows
hasPart
used the previous fifteen words and next 25 words -excluding named entities- as their bag of words
created a 'bag of words' for each person in our dataset, comprised of all words that appear before and after the person's name in the ONDB
hasPart
hasObjective
generate our topic model
hasPart
follows
removed all named-entity mentions in these biographies
hasPart
hasPart
hasPart
follows
dropped words that are in a standard stop list
converted the remaining words into lowercase
follows
applied the Porter stemmer
follows
hasObjective
strip away standard English suffixes in a specific order

SO-driven knowledge extraction from text

We used Random Forests in order to perform the classification experiment and then we evaluated the results .

In addition, we conducted two more experiments using SVM and Logistic Regression respectively .

identify textual chunks

We used Random Forests in order to perform the classification experiment and then we evaluated the results .

In addition, we conducted two more experiments using SVM and Logistic Regression respectively .

extract entities

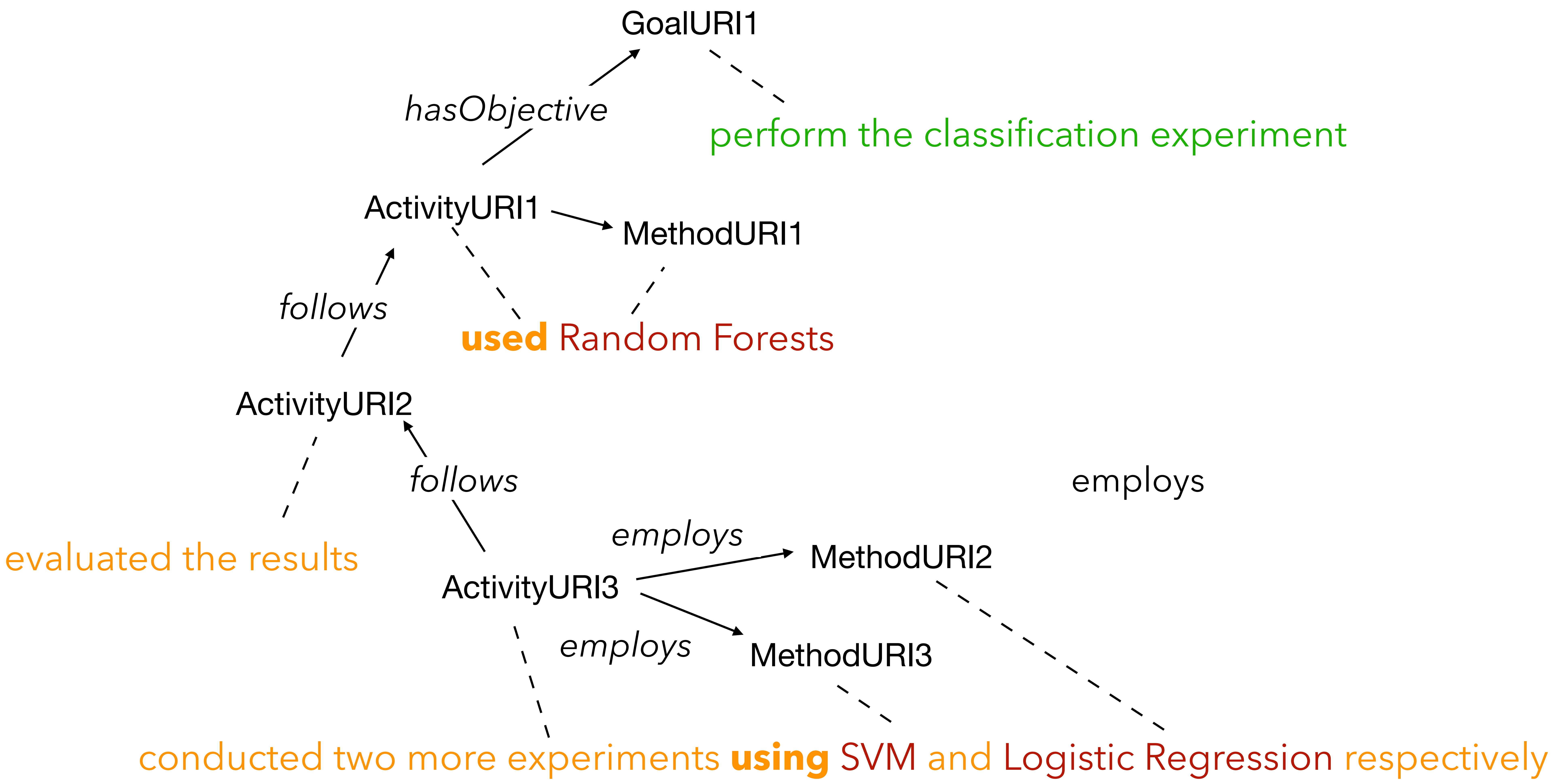
Activity Method Goal Activity

We used Random Forests in order to perform the classification experiment and then we evaluated the results

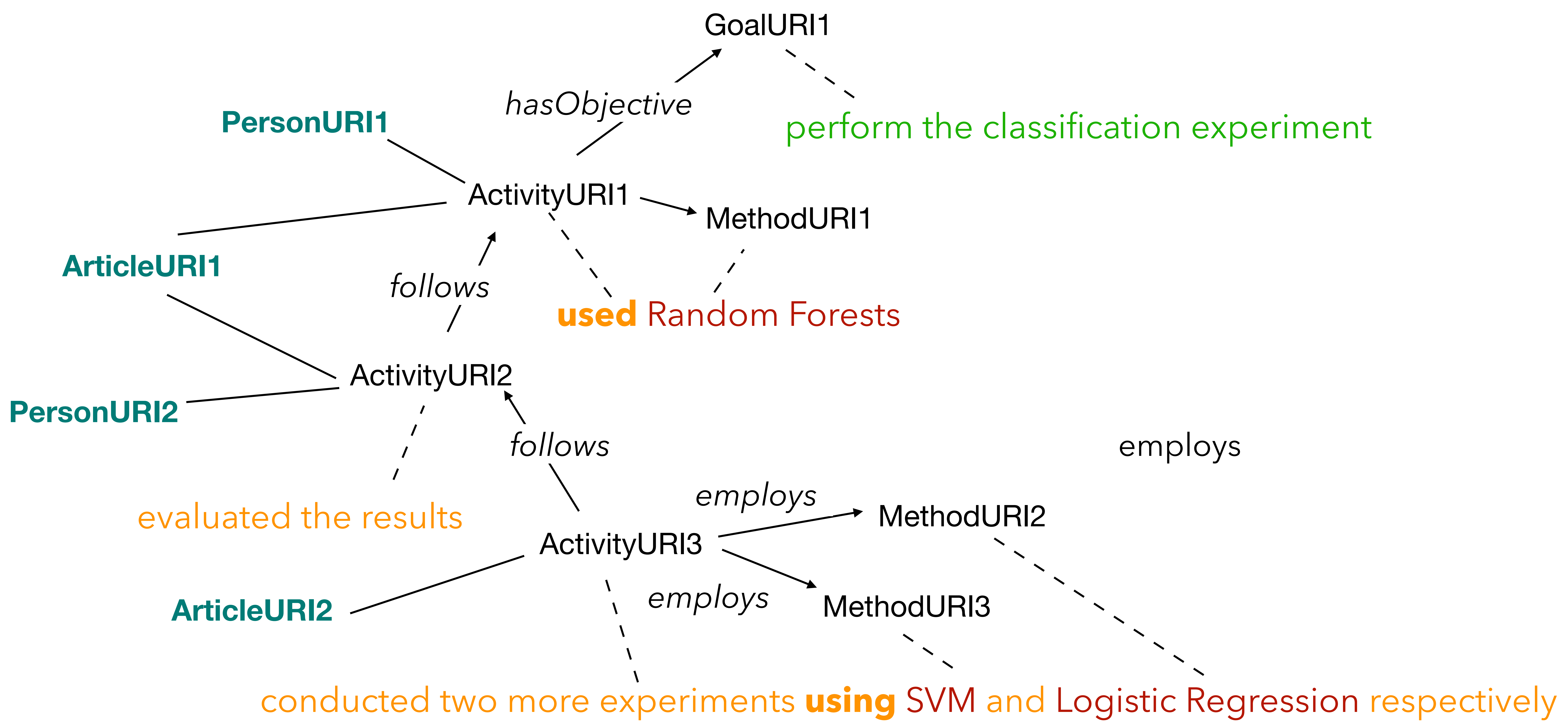
Activity Method Method

In addition, we conducted two more experiments using SVM and Logistic Regression respectively

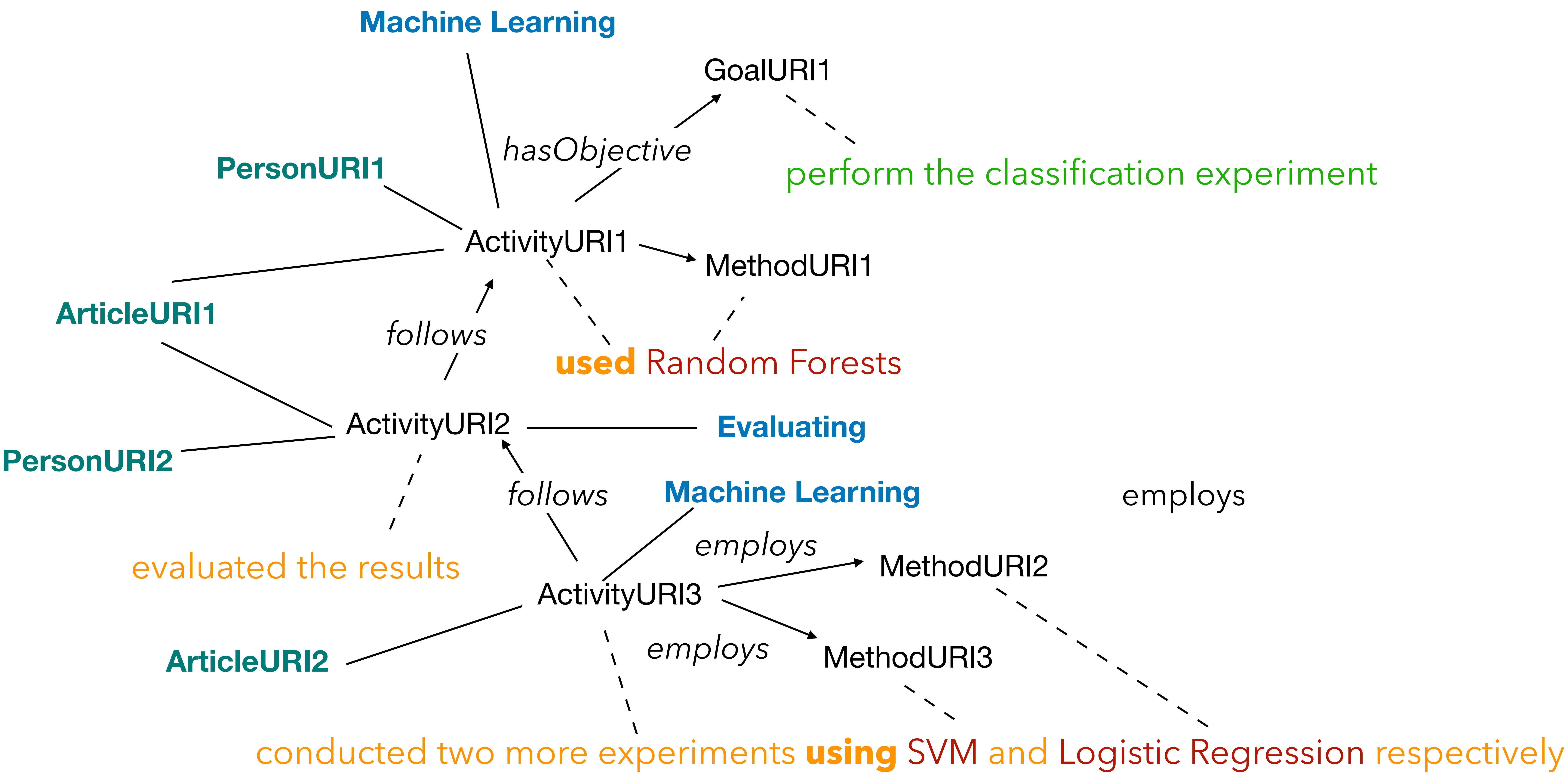
extract relations



add actor and resource metadata



assign activity types



link with existing knowledge

semantic paths through common Activity Types

semantic paths through common Actors

Evaluation:

Rule-based extraction

Source: 50 research articles from Digital Humanities, Geology, Medicine, Bioinformatics, Biology, Computer Science, Sociology and Anthropology.

Reference standard:

- Annotations by two human annotators (inter-annotator agreement: 81% kappa)
- Duration for manual annotation: 3.5 - 4 hrs per article!

Dataset:

- 1700 Activities, 300 Goals, 700 Propositions,
- 1000 follows(), 100 hasPart(), 250 hasObjective(), 200 resultsIn(), 400 employs()

Methodology:

- token-based, entity-based evaluation (threshold: 86%)
- micro- & macro-averaging

Evaluation

Rule-based extraction

Results:

Entity Evaluation:

	Entity-based	Token-based
Entity Type	**F1**	**F1**
Activity	0,72	0,81
Goal	0,76	0,80
Proposition	0,79	0,82
Method	0,91	0,85

Relation Evaluation:

Relation Type	F1
follows	0,71
hasPart	0,55
hasObjective	0,79
resultsIn	0,56
employs	0,90

Error sources:

- Human errors (author / editor misspellings)
- External modules
- Rules and constraints

Using machine learning techniques:

- Algorithms used: Logistic regression, SVM, Random forests
- Combinations of handcrafted features and word embeddings
- Pipeline proposed splitting sentence from token classification

Extracting **Activity** and **follows(Activity, Activity)**

Activity

F1 scores

Eval method	**Test sets**			
	DH	BIOINF	MED	ALL
Token-based	0,82	0,88	0,92	0,88
Entity-based	0,73	0,72	0,71	0,72

follows(Activity, Activity)

F1 scores

Test sets			
DH	BIOINF	MED	ALL
0,87	0,86	0,92	0,89

Use cases:

- Find information on earlier work relevant to one's own research
- Goal-oriented organization of research work and project planning
- Discovery of connections between resources, tools and methods
- Gathering evidence of the use of digital resources for scholarship
- Critical evaluation of digital humanities

Acknowledgements:

Special thanks go to Ion Androutsopoulos for guidance with NLP methods

Thanks also go to Stavros Angelis, Agiatis Benardou, Costis Dallas, Lorna Hughes, Leonidas Papachristopoulos, Eliza Papaki for contributions and insights during the ontology development

References:

V. Pertsas and P. Constantopoulos, “Scholarly Ontology: modelling scholarly practices”, Intl Journal on Digital Libraries, Vol. 18 (3), pp. 173–190. 2017.

V. Pertsas and P. Constantopoulos, “Ontology-Driven Information Extraction from Research Publications”, Proc. 22nd International Conference on Theory and Practice of Digital Libraries, TPDL 2018, Porto, Sept. 2018.

V. Pertsas, P. Constantopoulos and I. Androutsopoulos, “Ontology Driven Extraction of Research Processes”, Proc. 17th International Semantic Web Conference, ISWC 2018, Monterey, CA., Oct. 2018.

L. Hughes, P. Constantopoulos, and C. Dallas, "Digital Methods in the Humanities: Understanding and Describing their Use across the Disciplines", in Susan Schreibman, Ray Siemens, John Unsworth (eds.), A New Companion to Digital Humanities, Wiley-Blackwell, 2016.

Thank you!

panosc@aueb.gr