From publications to knowledge graphs

Panos Constantopoulos and Vayianos Pertsas

Athens University of Economics and Business & Athena Research Centre

WHO EIOS Global Technical Meeting, Seoul, 12-14 Nov. 2019

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?

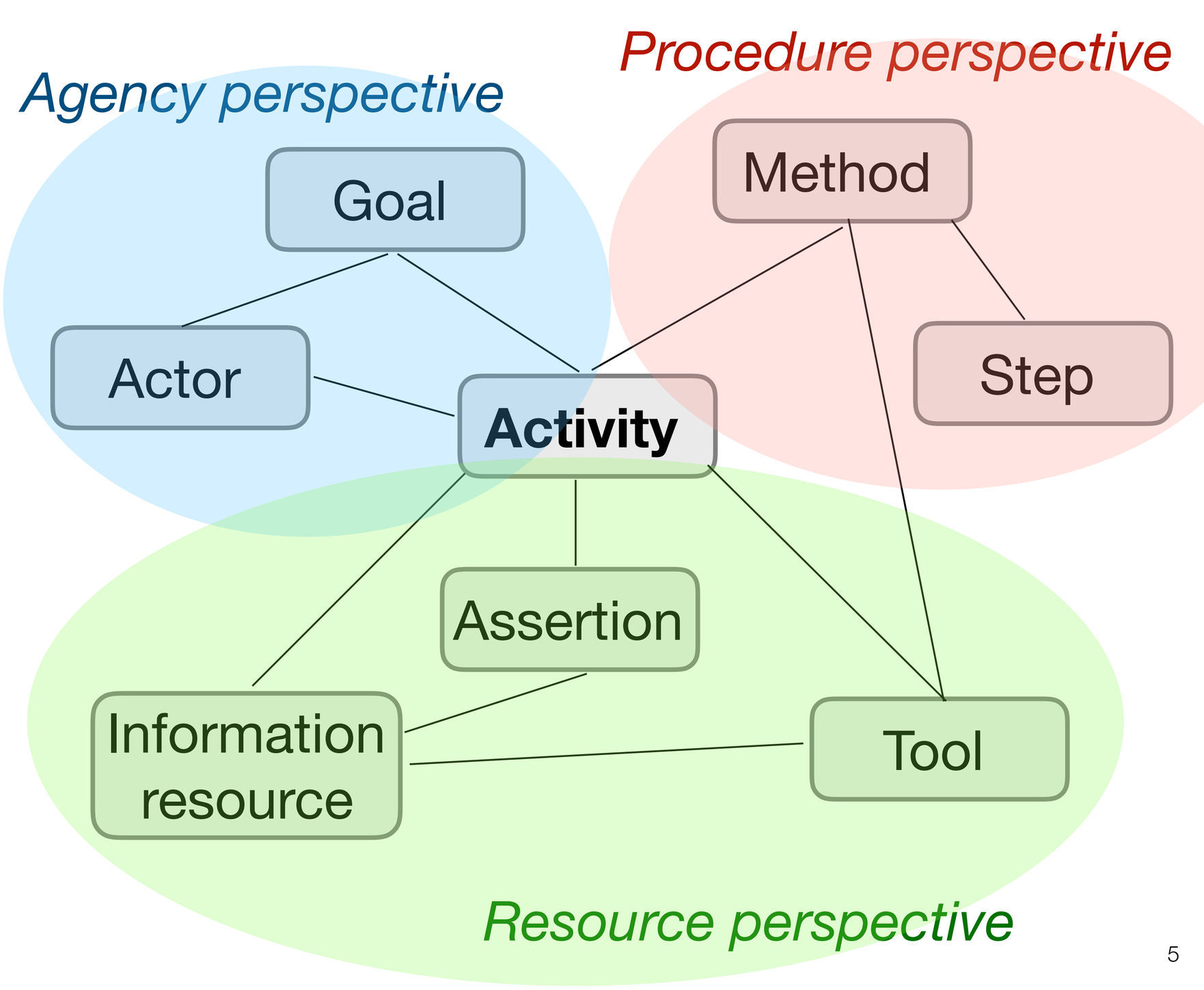
Answering these questions today:

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

Representing scholarly work

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



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

The NeDiMAH Methods Ontology - NeMO

www.nedimah.eu

An ontology for Digital Humanities research, including classification and a shared vocabulary

Incorporates previous DH taxonomies (Oxford ICT, TaDiRAH, DHCommons)

CIDOC CRM compatible

Network for Digital Methods in the Arts and Humanities - NeDiMAH
(ESF 2011-15)

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.

Why use an ontology?

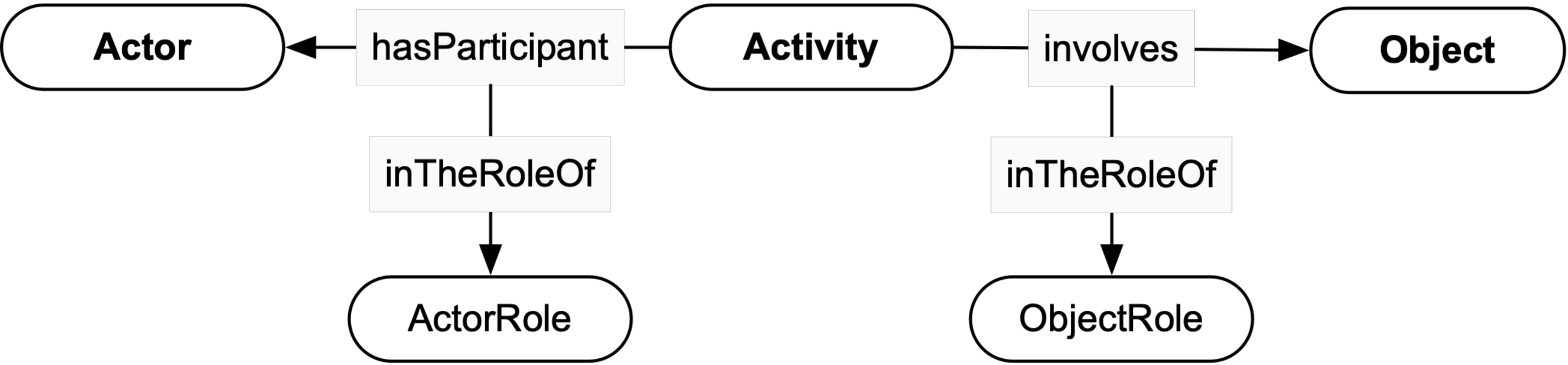
Provides a formalization of basic concepts.

Provides a conceptual framework for complex query answering.

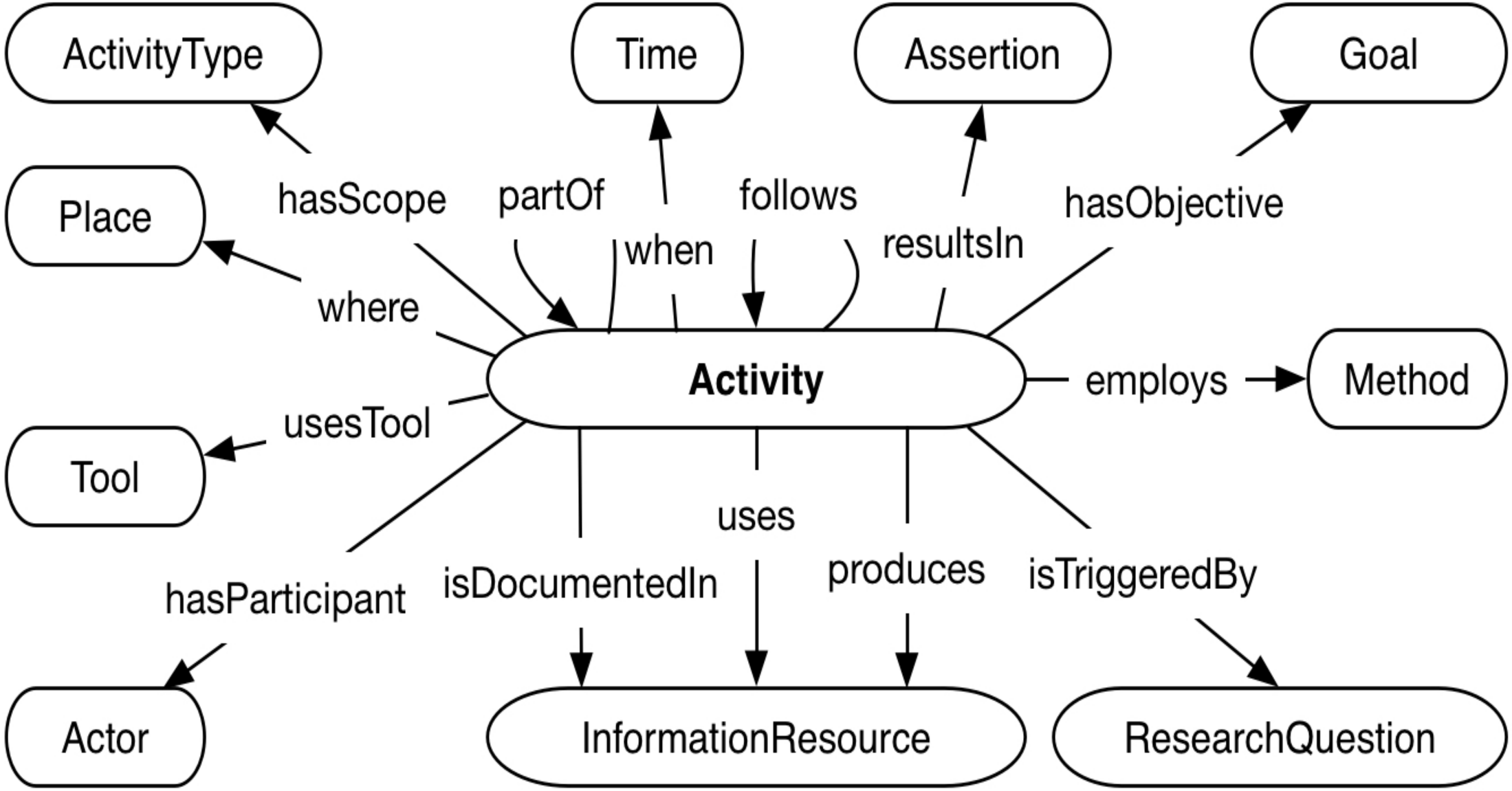
Acts as semantic glue between different taxonomies.

Supports the development of an ecosystem of interoperable resources and services for discovering, understanding, selecting, linking and contributing content, tools and methods.

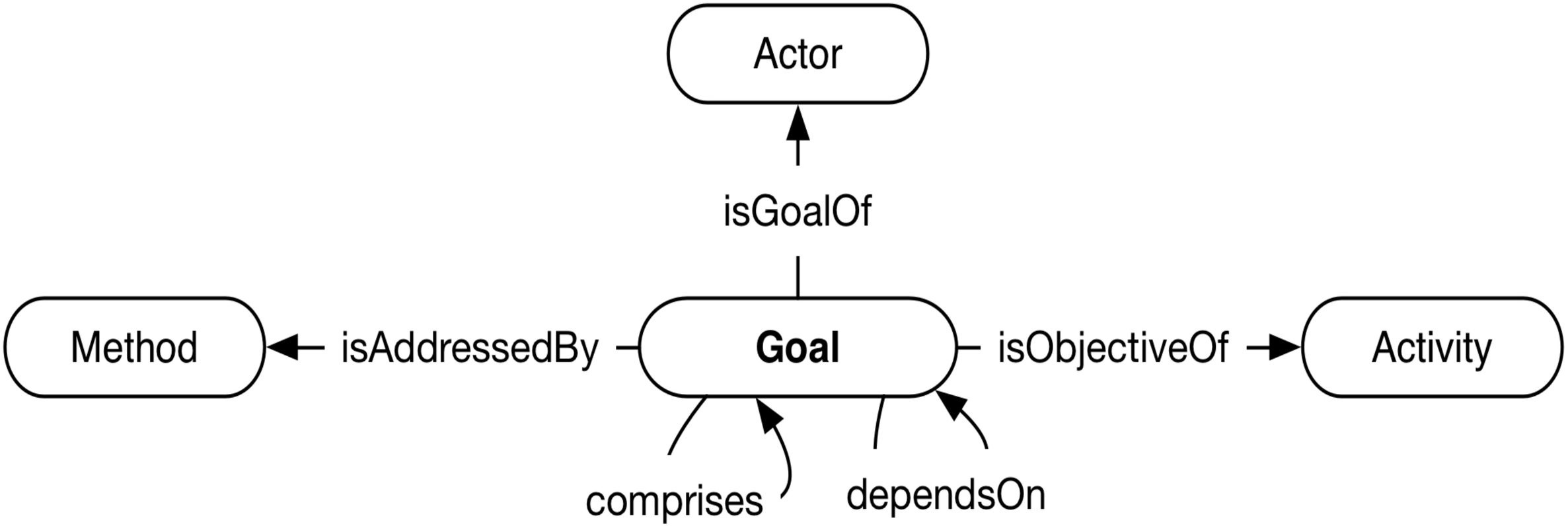
SO: Top Elements



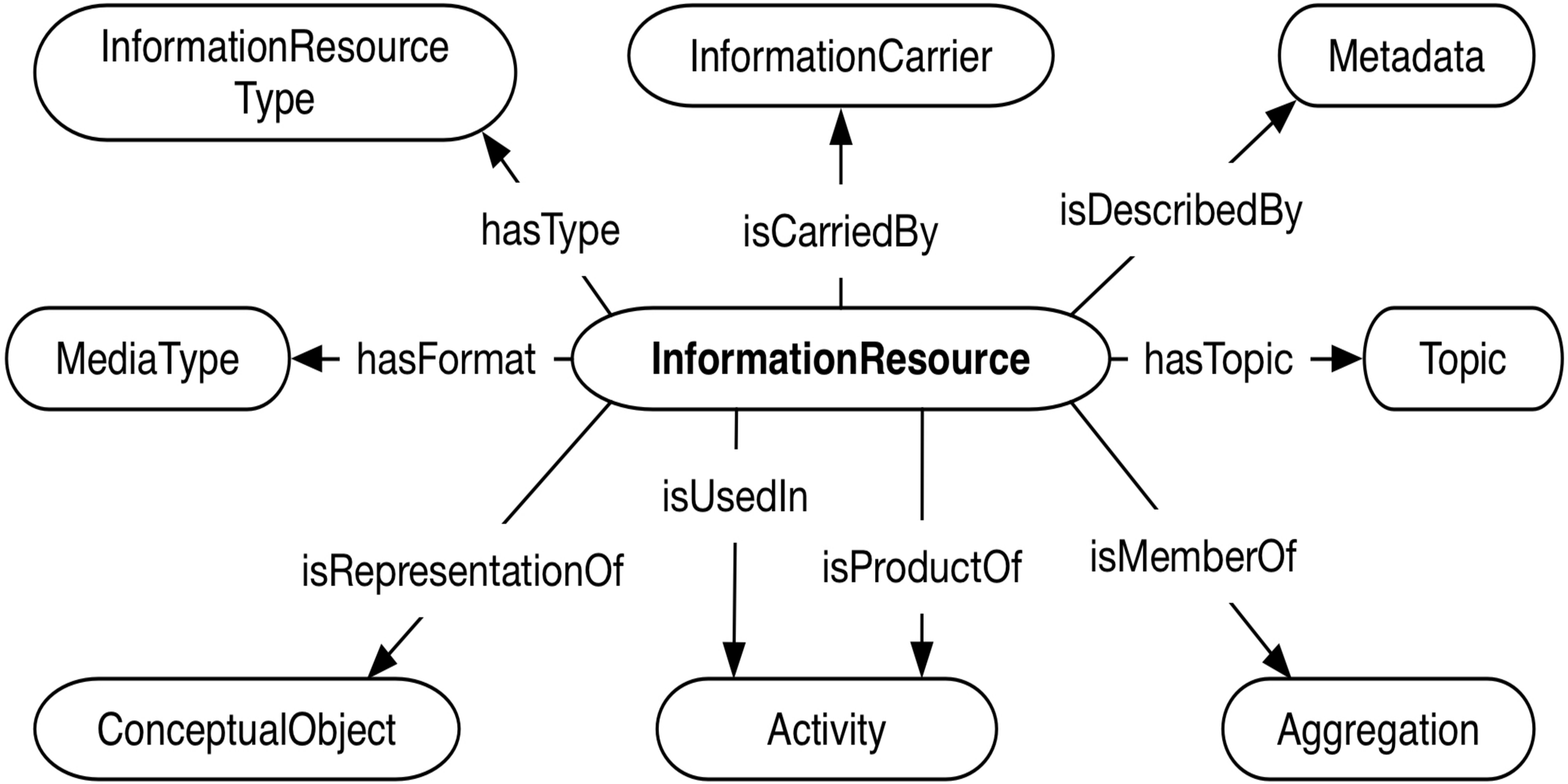
SO: Activity Perspective



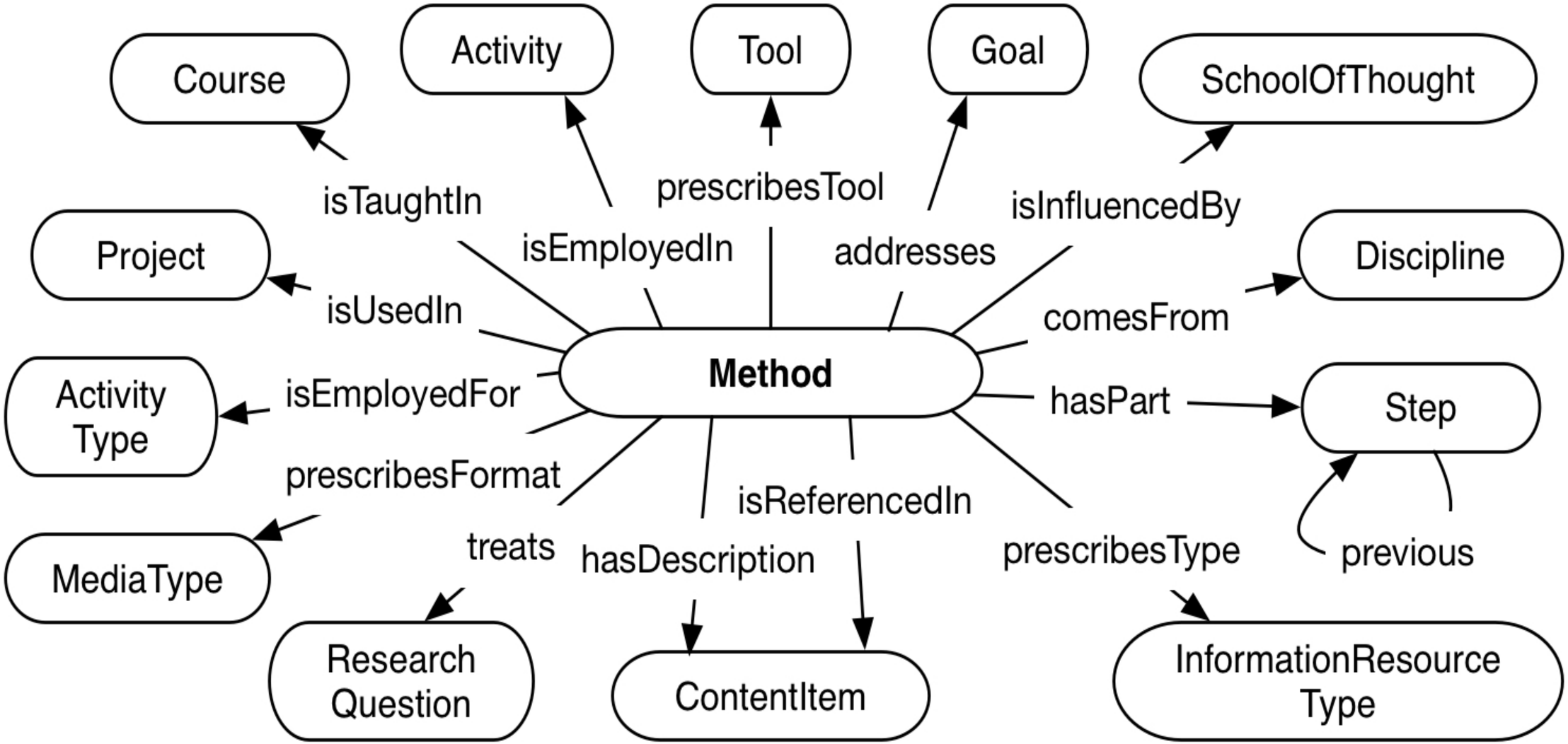
SO: Agency Perspective



SO: Resource Perspective

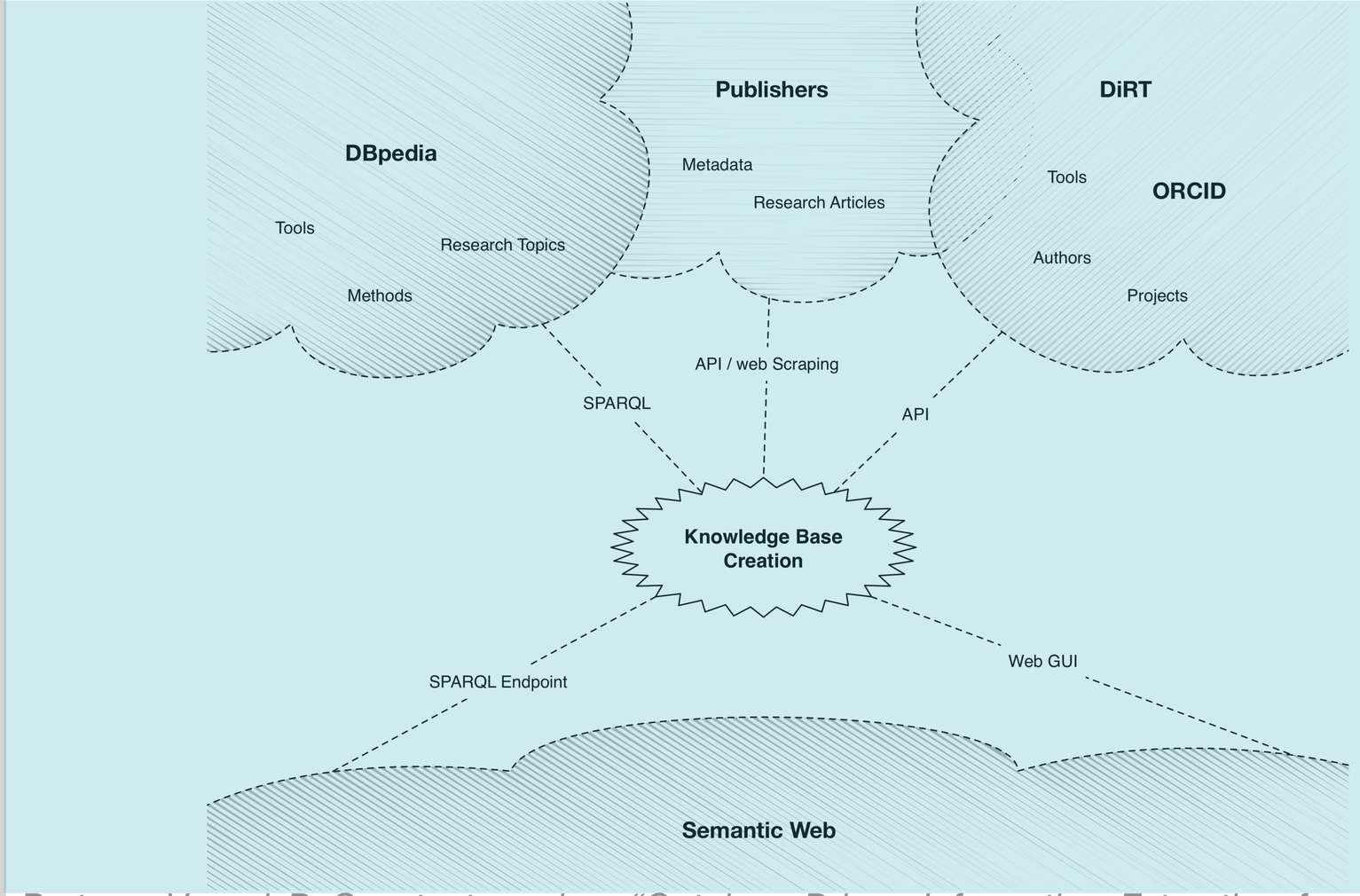


SO: Procedure Perspective



Knowledge base creation

Research Spotlight



Pertsas, V. and P. Constantopoulos, "Ontology-Driven Information Extraction from Research Publications", Proc. 22nd International Conference on Theory and Practice of Digital Libraries, TPDL 2018, Porto, Portugal, Sept. 2018, LNCS 11057, pp. 241–253. https://doi.org/10.1007/978-3-030-00066-0_21 .

Pertsas, V., P. Constantopoulos and I. Androutsopoulos, "Ontology-Driven Extraction of Research Processes", Proc. 17th International Semantic Web Conference Monterey, CA, USA, Oct. 2018, ISWC 2018, LNCS 11136, pp. 162–178. https://doi.org/10.1007/978-3-030-00671-6_10 . Spotlight paper, nominee for best student paper award.

Information extraction from publications

Challenges

Exploiting metadata

Named entities of non-common type (e.g. research methods)

Non-named entities (such as activities, goals, propositions)

Semantic association of extracted entities

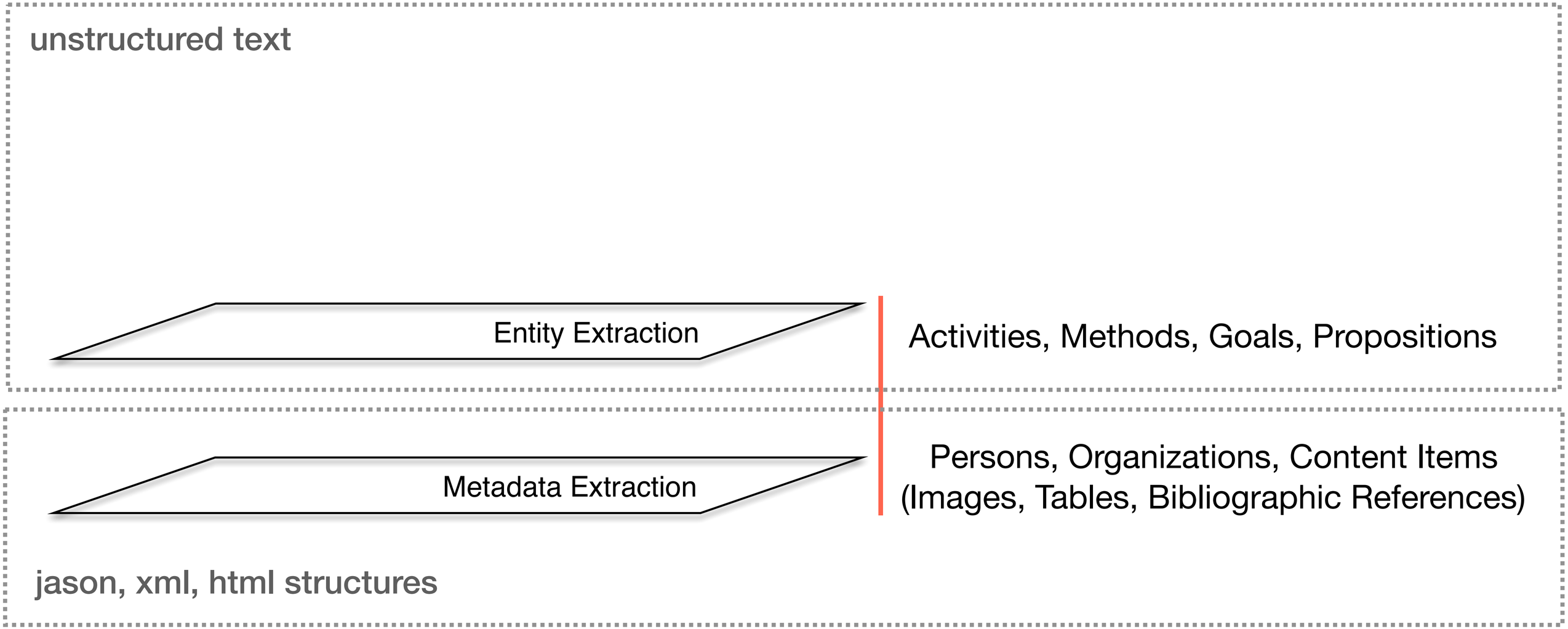
Linking with other published information

Knowledge base creation - Stage 1

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

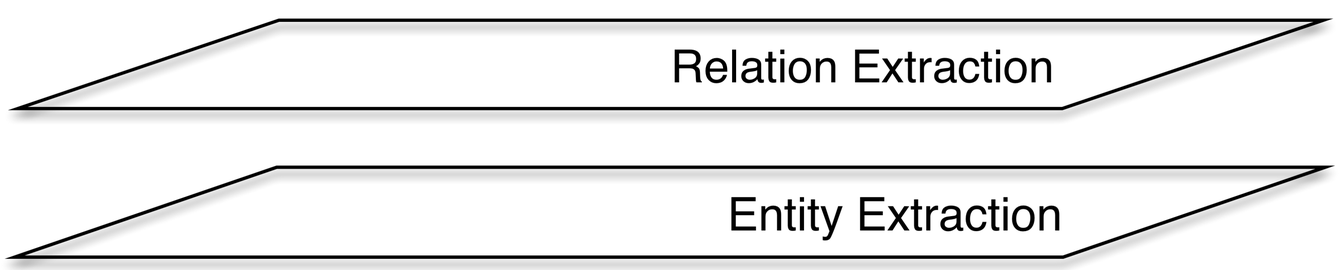
jason, xml, html structures

Knowledge base creation - Stage 2



Knowledge base creation - Stage 3

unstructured text



follows, hasPart, hasObjective, employs, resultsIn, hasParticipant, hasTopic, etc.

Activities, Methods, Goals, Propositions

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

jason, xml, html structures

Knowledge base creation - Stage 4

unstructured text

URI Creation - Linked Data Publishing

owl:sameAs, owl:equivalentProperty, rdfs:Label, skos:altLabel

Relation Extraction

follows, hasPart, hasObjective, employs, resultsIn, hasParticipant, hasTopic, etc.

Entity Extraction

Activities, Methods, Goals, Propositions

Metadata Extraction

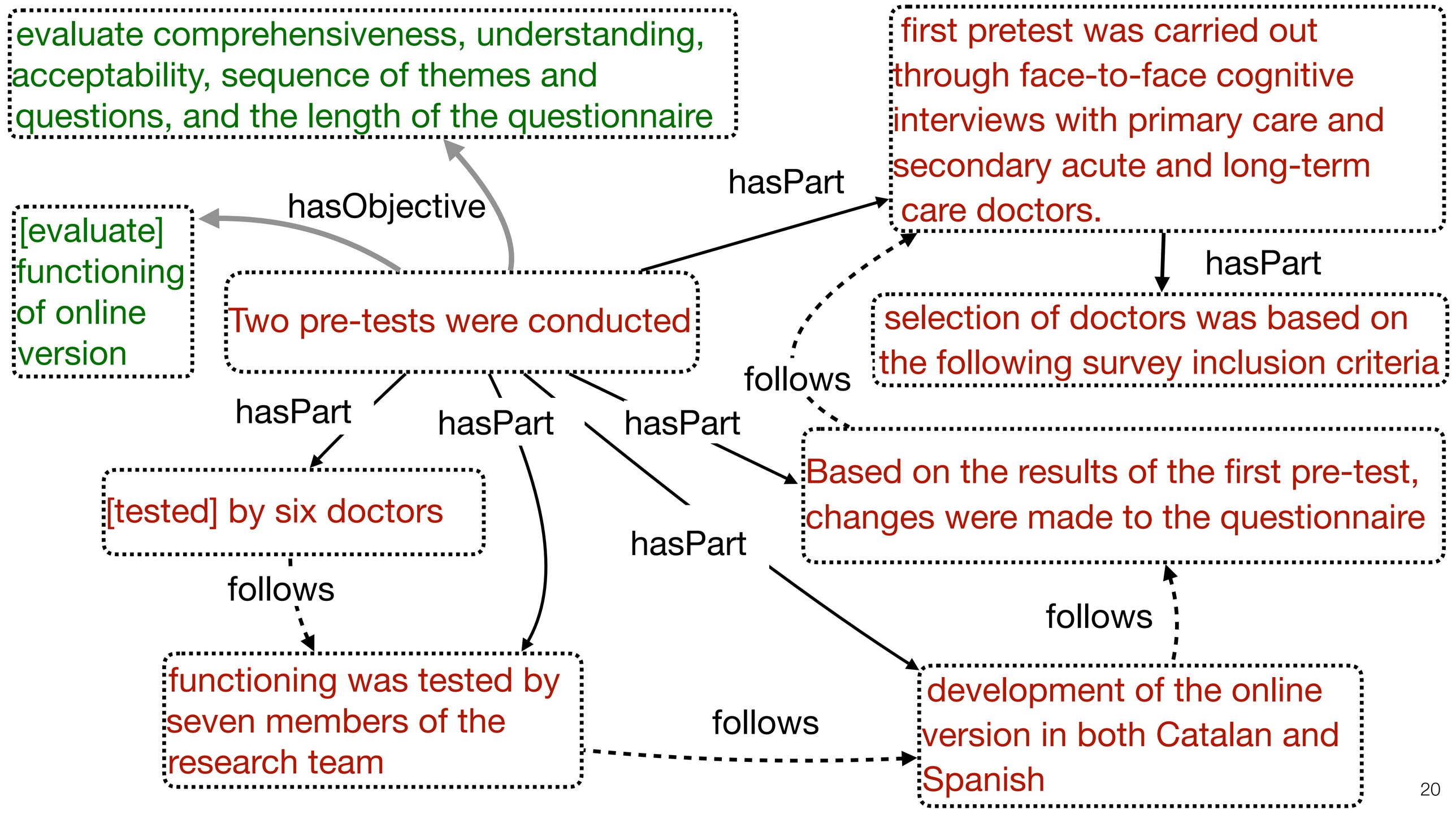
Persons, Organizations, Content Items (Images, Tables, Bibliographic References)

jason, xml, html structures

Representing activity structure

From: María-Luisa Vázquez Navarrete et al., Adapting the COORDENA questionnaire for measuring clinical coordination across health care levels in the public health system of Catalonia (Spain), Public Health Panorama, 4(4), Dec. 2018, 491–735.

Two pre-tests were conducted in order to evaluate: firstly, comprehensiveness, understanding, acceptability, sequence of themes and questions, and the length of the questionnaire; and secondly, the functioning of the online version. The first pretest was carried out through face-to-face cognitive interviews with primary care and secondary acute and long-term care doctors. The selection of doctors was based on the following survey inclusion criteria: doctors had worked for at least one year in the health care organization, doctors provided direct care to patients, and their daily practice involved contact with doctors from other care levels. Based on the results of the first pre-test, changes were made to the questionnaire, followed by the development of the online version in both Catalan and Spanish. Its functioning was tested first by seven members of the research team and then by six doctors.


evaluate comprehensiveness, understanding, acceptability, sequence of themes and questions, and the length of the questionnaire
hasObjective
[evaluate] functioning of online version
Two pre-tests were conducted
hasPart
first pretest was carried out through face-to-face cognitive interviews with primary care and secondary acute and long-term care doctors.
hasPart
selection of doctors was based on the following survey inclusion criteria
follows
hasPart
hasPart
hasPart
[tested] by six doctors
follows
functioning was tested by seven members of the research team
Based on the results of the first pre-test, changes were made to the questionnaire
hasPart
follows
development of the online version in both Catalan and Spanish
follows

Knowledge extraction driven by the Scholarly Ontology

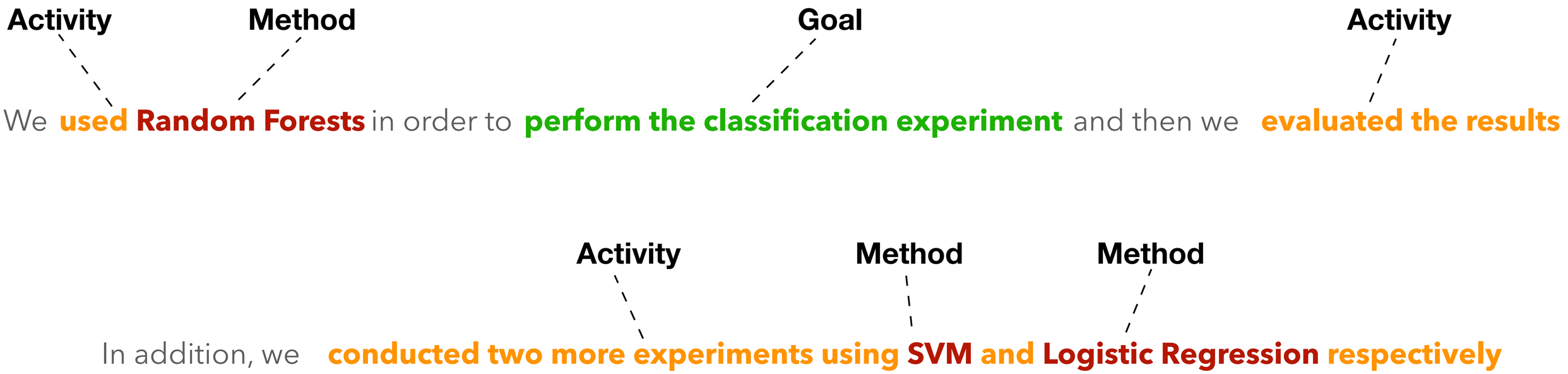
“We used Random Forests in order to perform the classification experiment and then we evaluated the results.
In addition, we conducted two more experiments using SVM and Logistic Regression respectively.”

1- Identify textual chunks

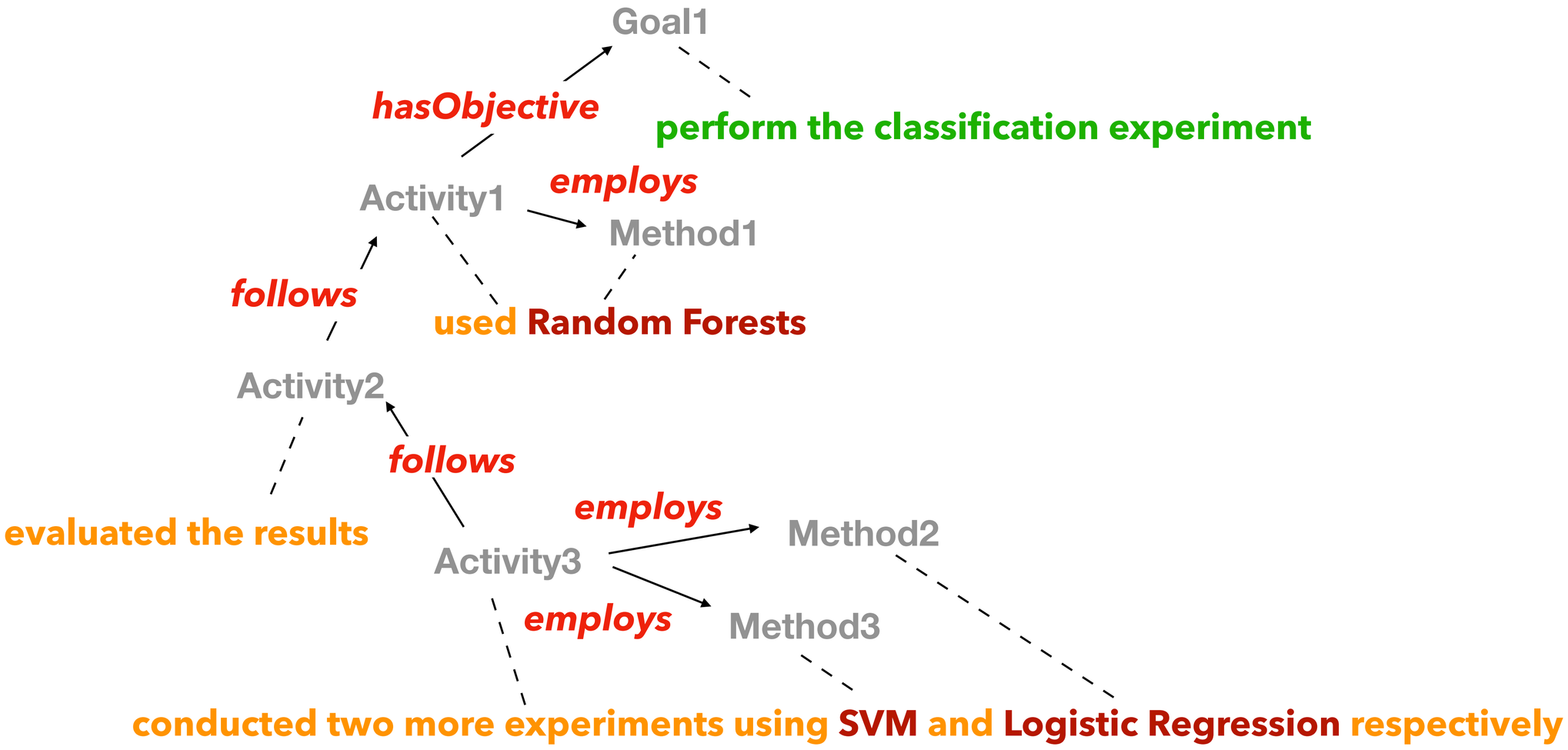
We used Random Forests in order to perform the classification experiment and then we evaluated the results.

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

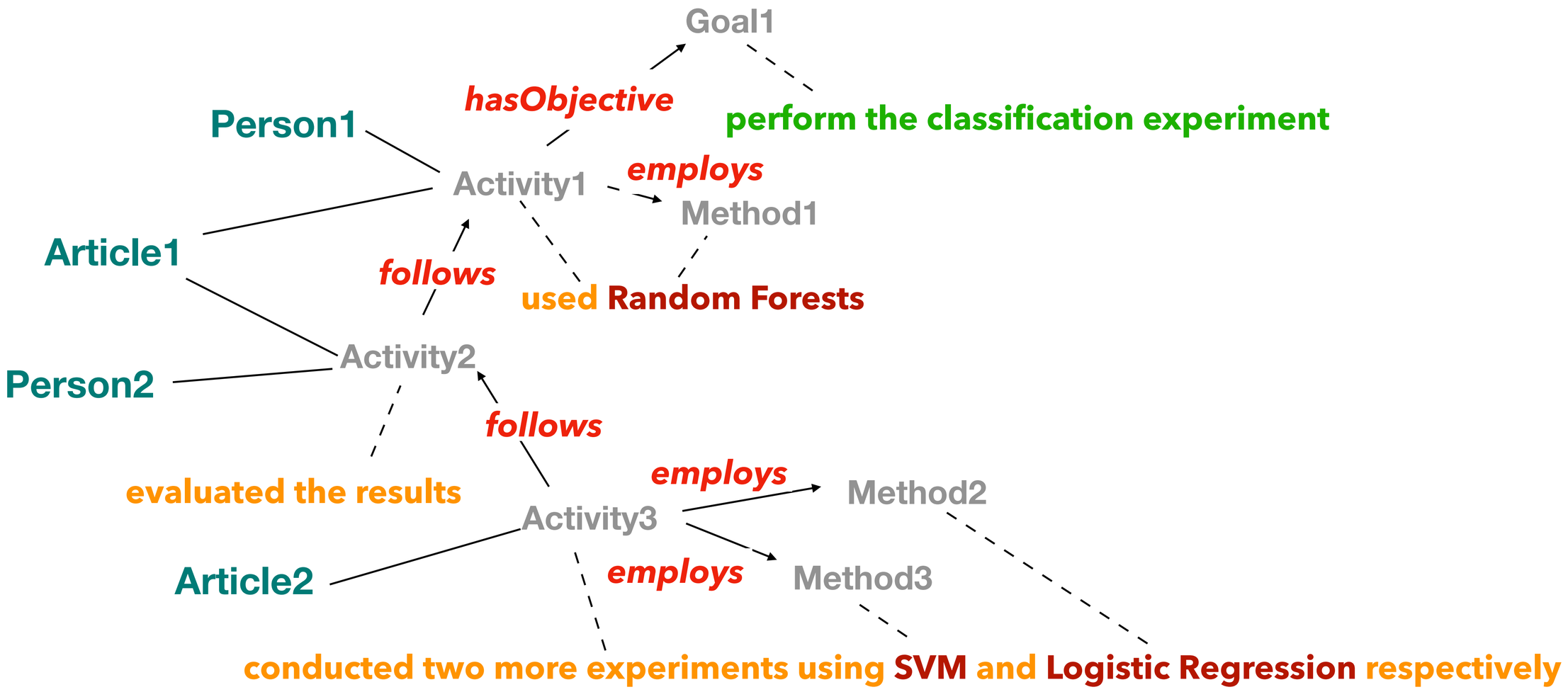
2- Extract entities



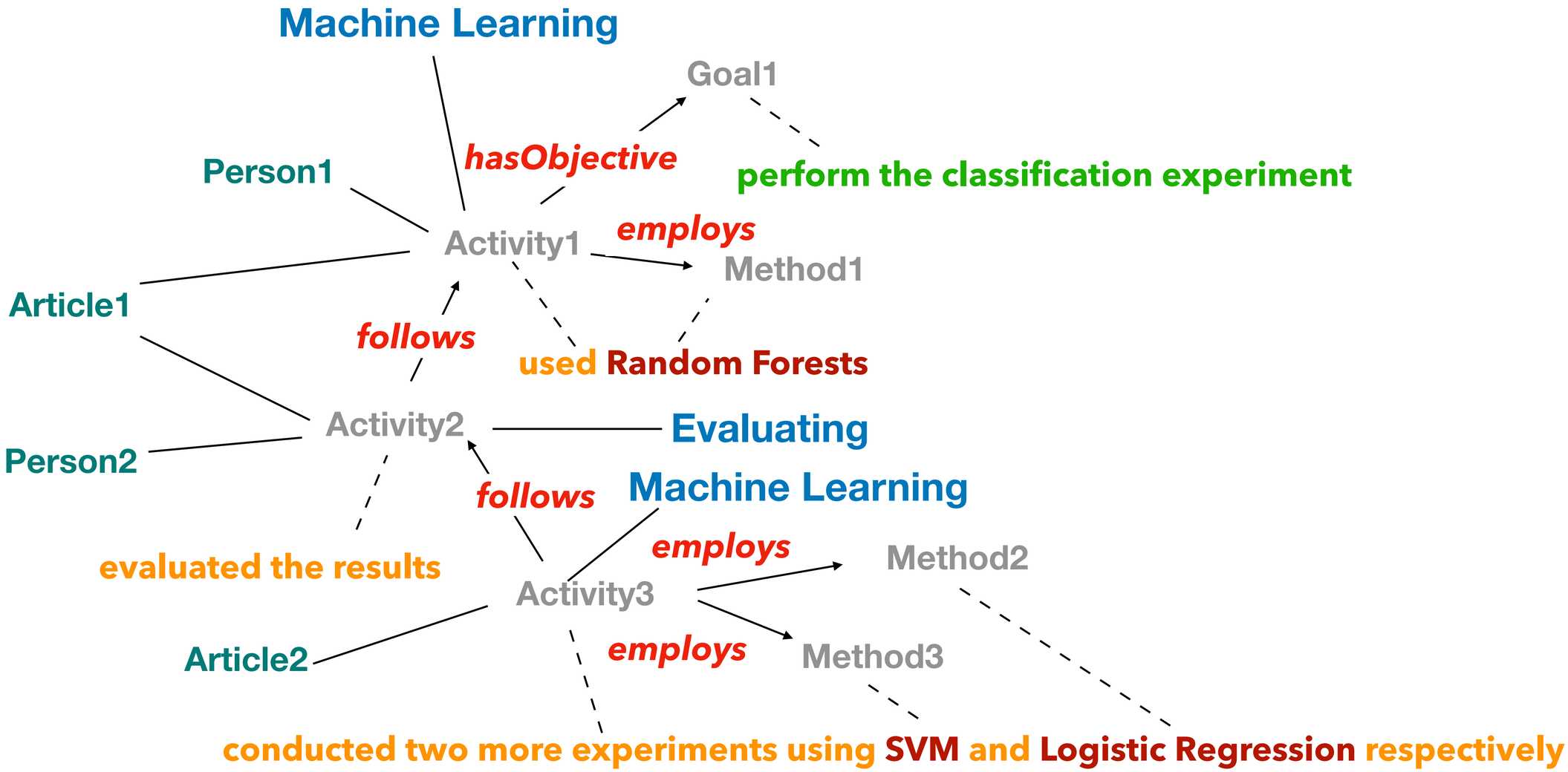
3- Extract relations



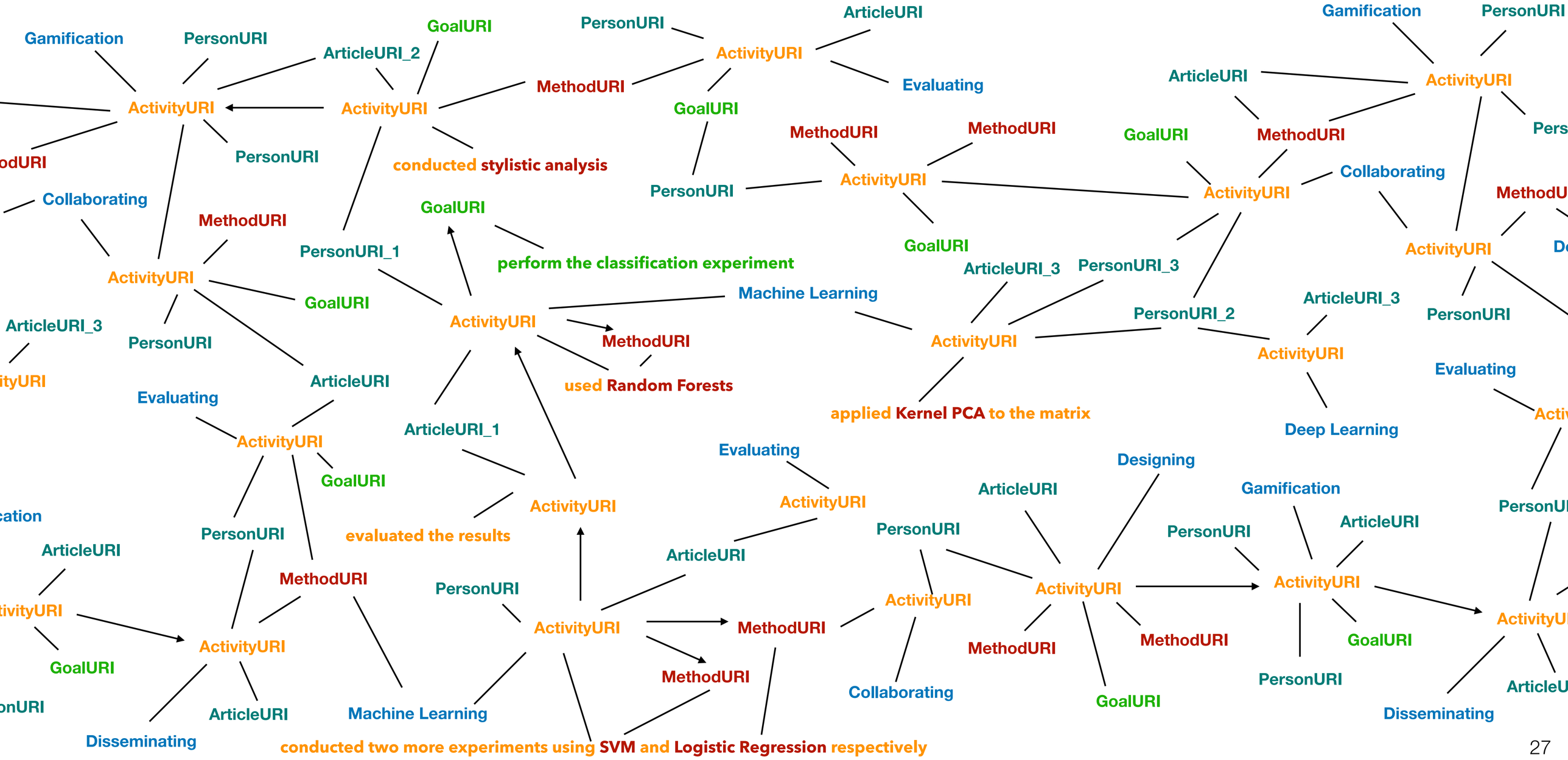
4- Add actor and resource metadata



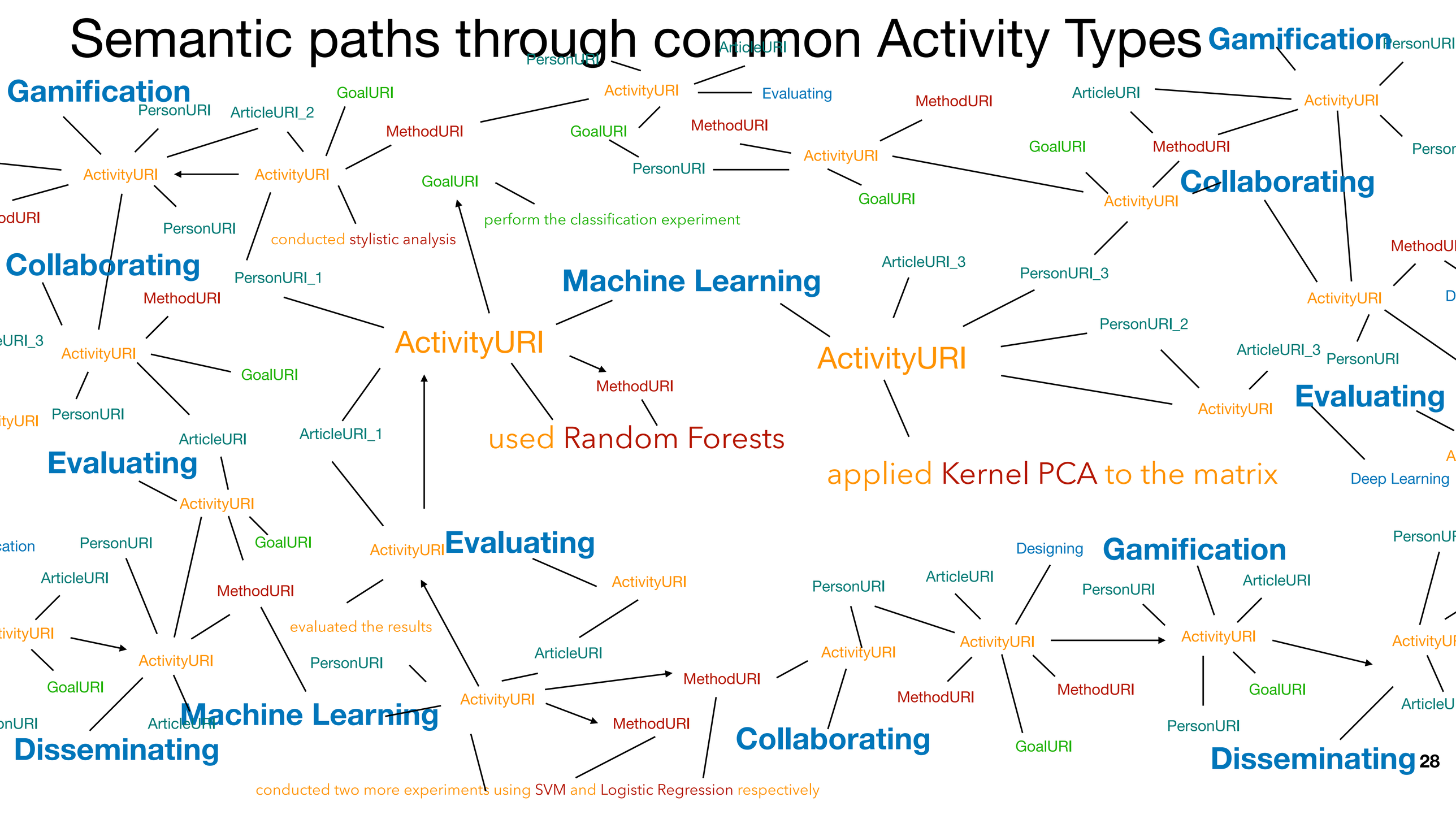
5- Assign activity types



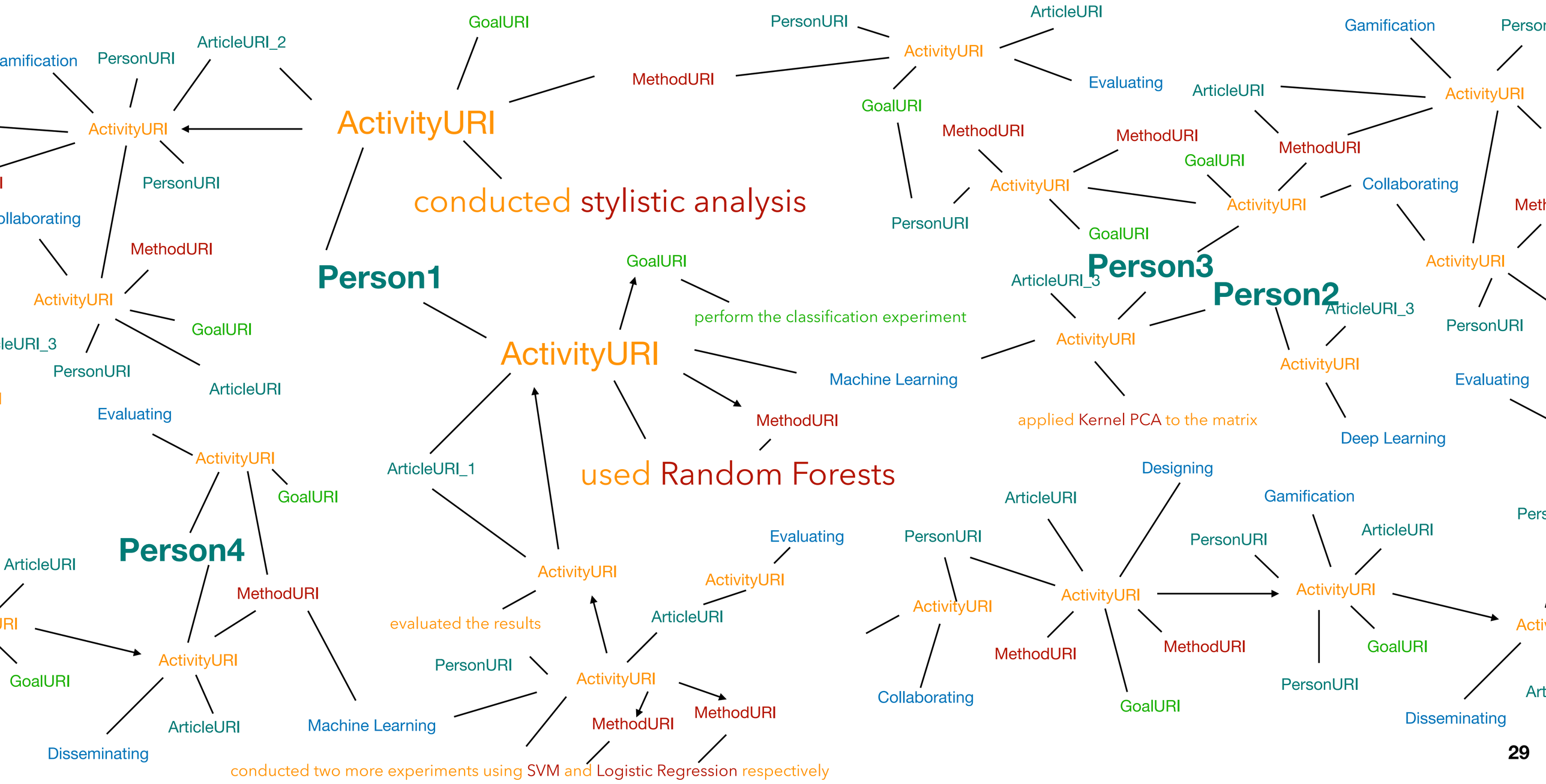
Link with existing knowledge



Semantic paths through common Activity Types



Semantic paths through common Actors



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

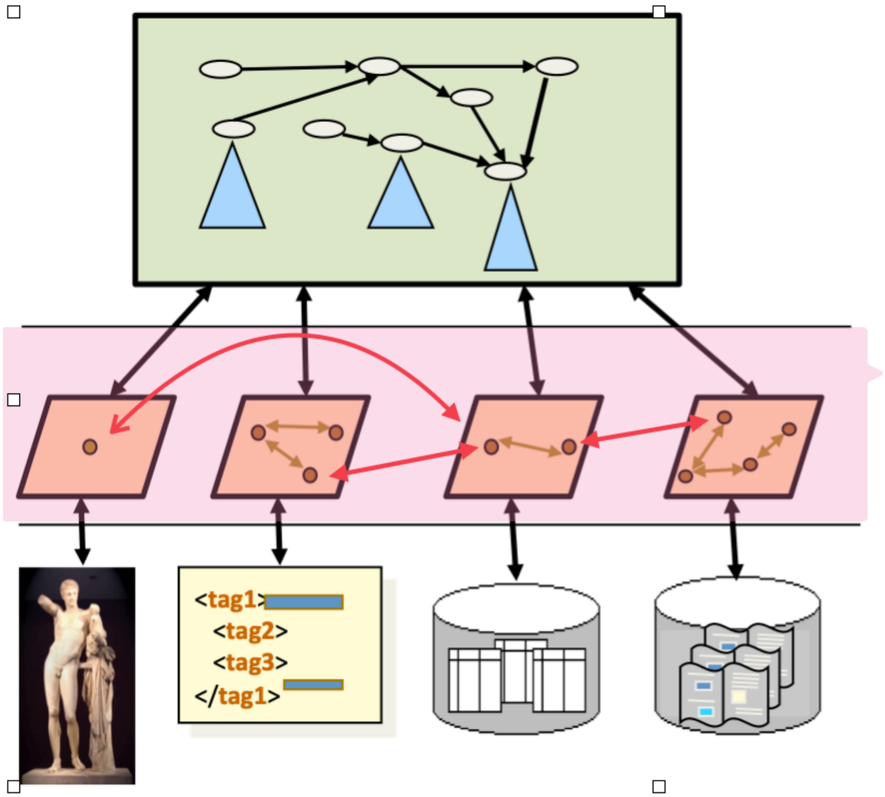
Gathering evidence of the use of resources

Mapping of research activity

Last one!

Virtual information spaces emerge around value-added semantic graphs that comprise assertions about objects AND processes:

what content items are about AND **why** and **how** people deal with them.



Acknowledgements:

Special thanks go to Ion Androutsopoulos for collaboration on 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 and to Vicky Dritsou for feedback and help with an example.

Thank you!

panosc@aueb.gr