Conference Linked Data: The ScholarlyData Project

Andrea Giovanni Nuzzolese[1] - Anna Lisa Gentile[2]
Valentina Presutti[1] - Aldo Gangemi[1,3] - Ziqi Zhang[4]

1. STLab, ISTC-CNR, Rome , Italy
2. IBM Research Almaden, San Jose, CA, USA
3. LIPN, Université Paris 13, Sorbone Cité, UMR CNRS, Paris, France
4. Nottingham Trent University, Nottingham, United Kingdom

May 28th 2017 - Portorož, Slovenia

Objectives

- Providing a sustainable and well grounded linked open dataset about scholarly facts

- Fostering a healthy growth of the dataset beyond the Semantic Web community

Background

- Since 2006 the Semantic Web community encourages the publication of Linked Data about scientific conferences in the field

- The Semantic Web Dog Food (SWDF)
  - reference linked dataset of the Semantic Web community
  - papers, people, organisations, and events
  - organised according to the Semantic Web Conference (SWC) ontology
  - collaborative data generation model

SWDF current status and motivations

- Lack of clear guidelines
- Use of vocabularies no longer maintained or existing
  - swrc-ext and xmllondon vocabularies
- Naïve usage of classes and properties
  - swc:room, swc:editorList, swc:completeGraph, swc:IW3C2Liaison
- Knowledge representation issues that prevent proper querying and reasoning
  - modelling of affiliations, roles and lists
- Duplicate entities

What we did

- Dataset refactoring from SWDF to ScholarlyData
  - best ontology design practices
  - permanent URIs
  - ontology alignment to other pertinent ontologies/vocabularies. e.g. SPAR, Organization Ontology, FOAF, SKOS
- Data cleansing and enhancement
  - entity deduplication
  - more linking with other datasets
- New data generation workflow

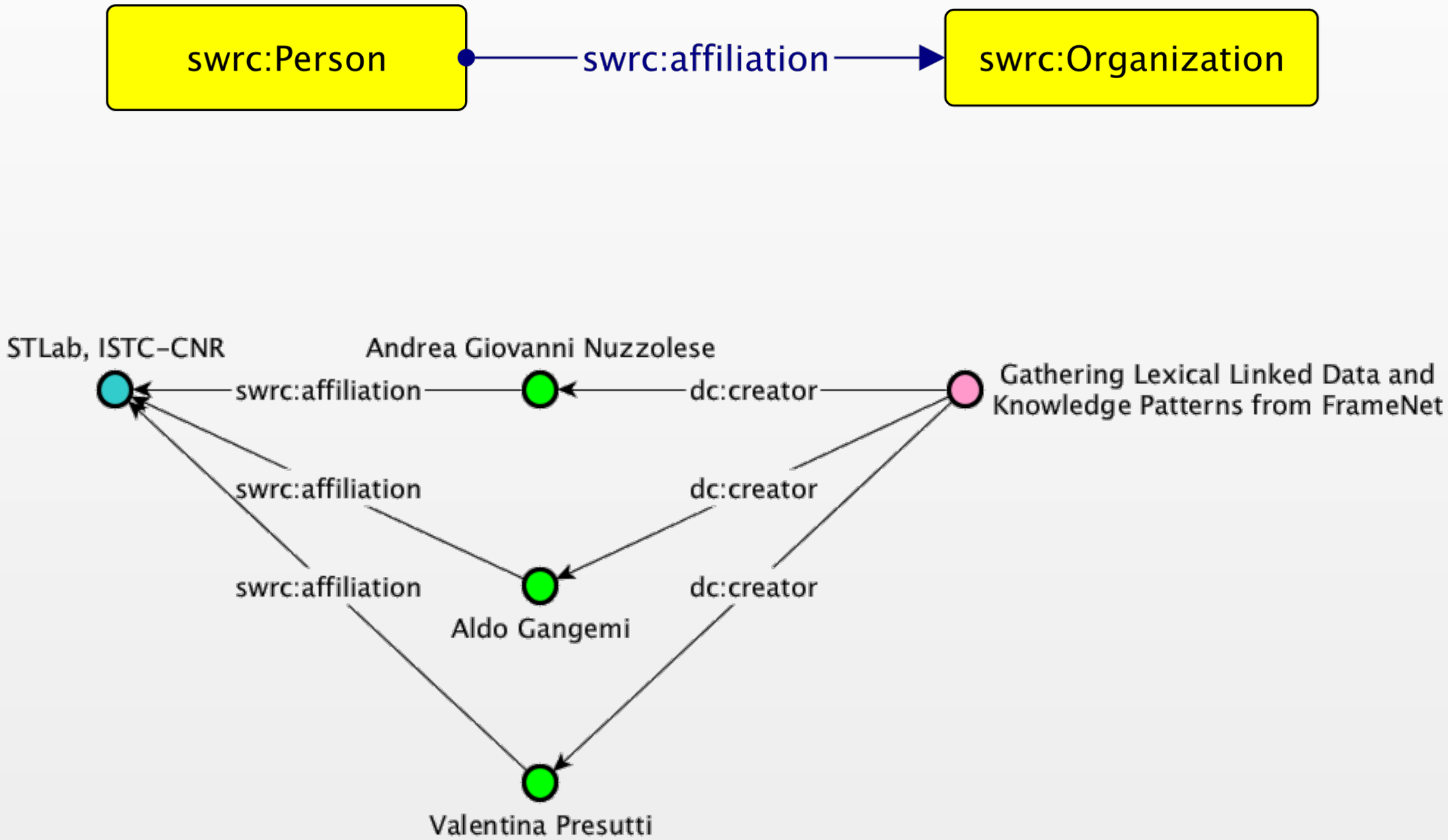
Before

ScholarlyData

After

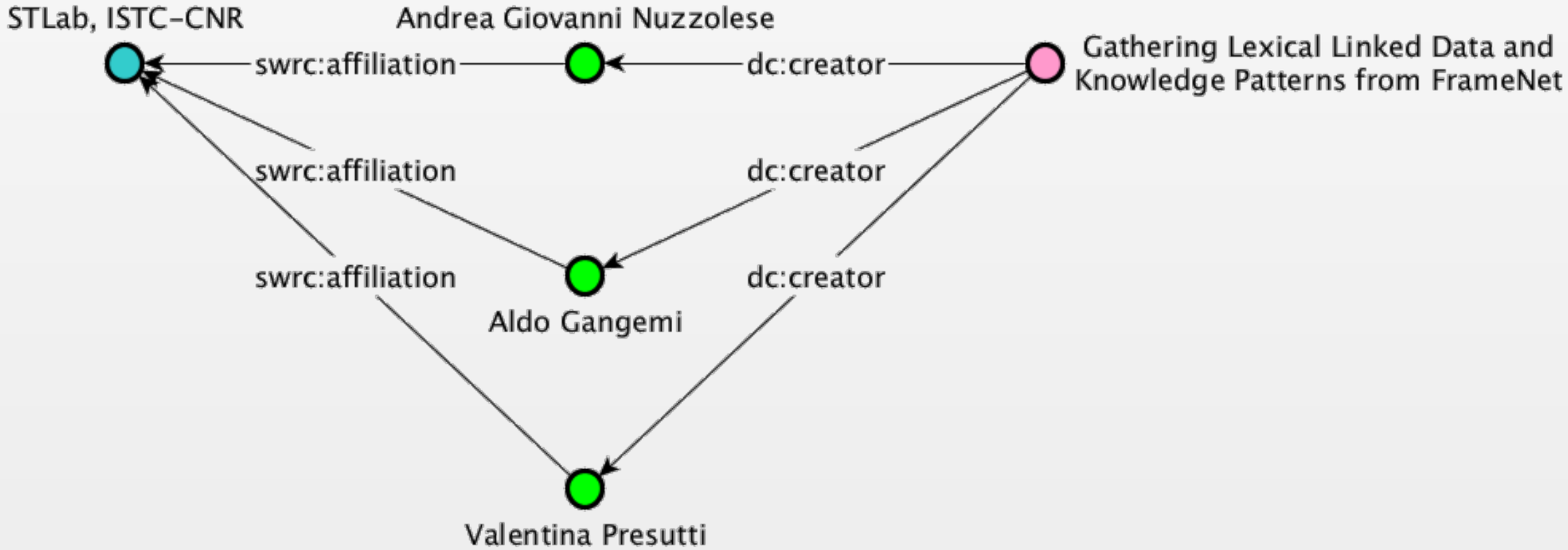
Affiliations: SWDF

Affiliations: SWDF



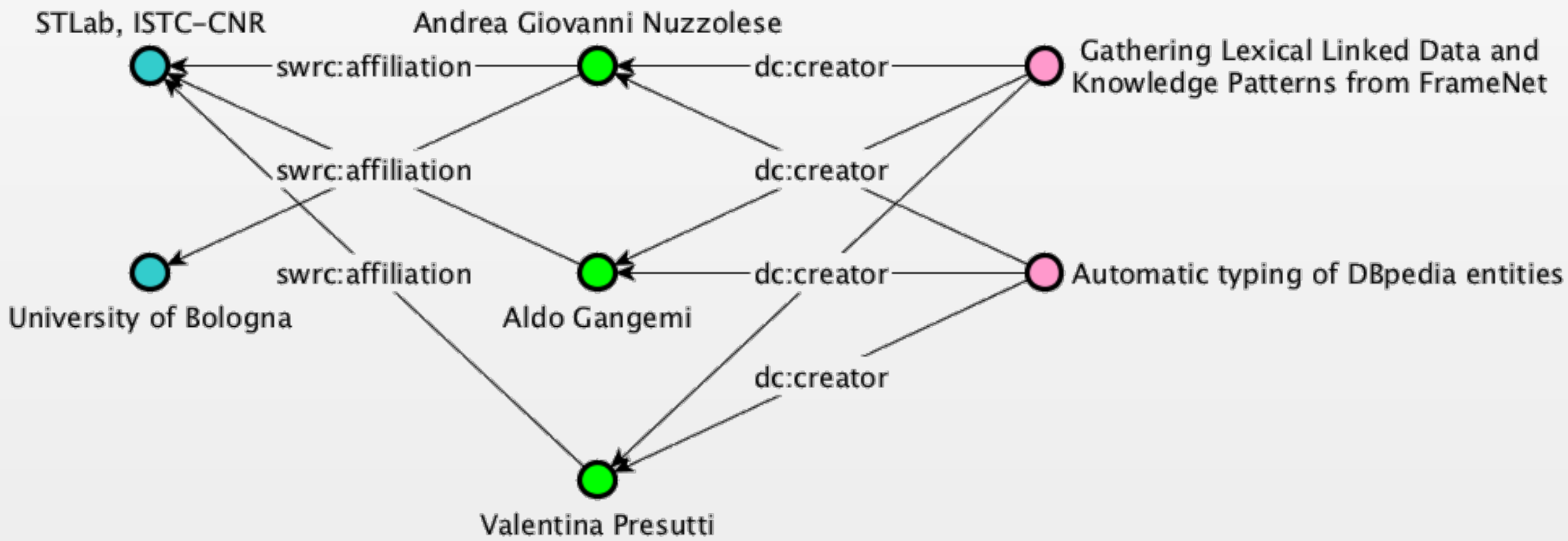
Affiliations: SWDF

What was the affiliation of Andrea Giovanni Nuzzolese when he authored the paper "Gathering Lexical Linked Data and Knowledge Patterns from FrameNet"?

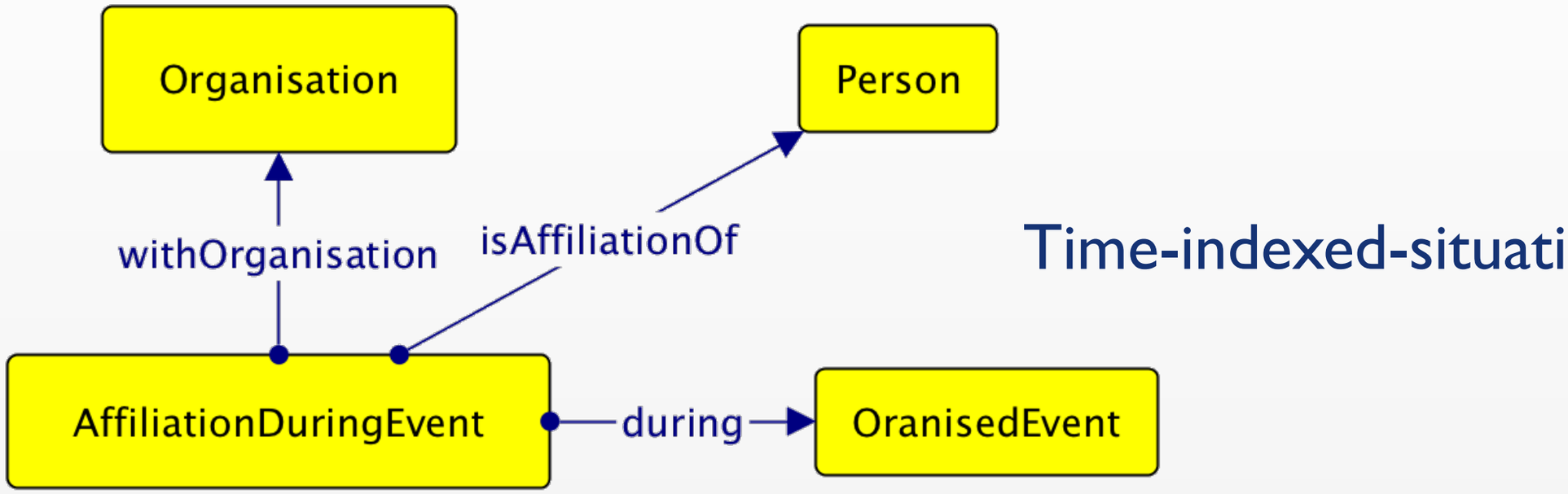


Affiliations: SWDF

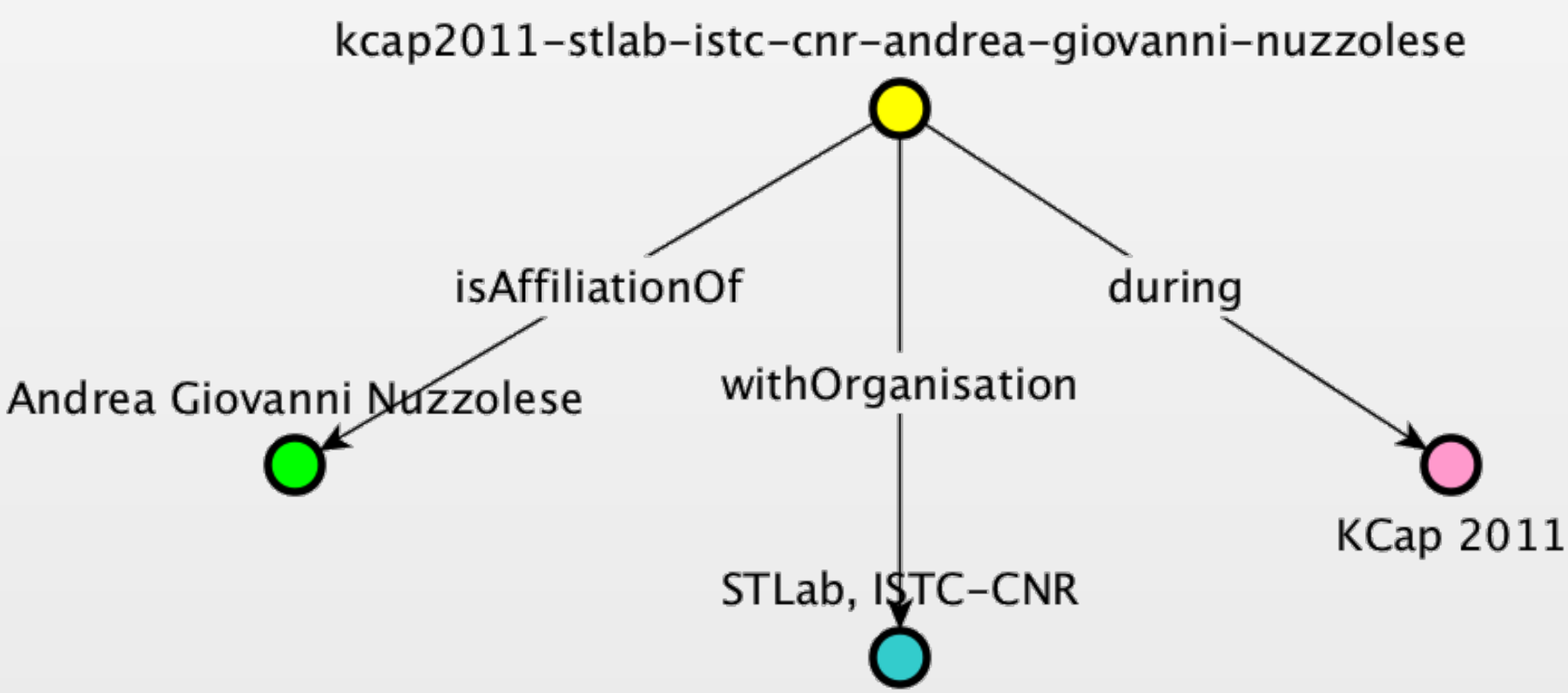
What was the affiliation of Andrea Giovanni Nuzzolese when he authored the paper "Gathering Lexical Linked Data and Knowledge Patterns from FrameNet"?



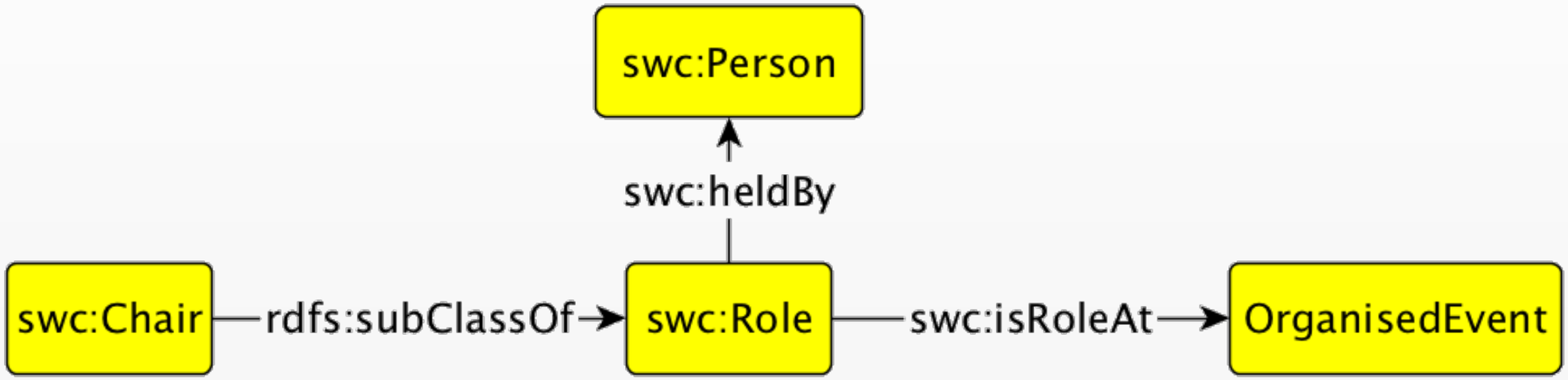
Affiliations: ScholarlyData



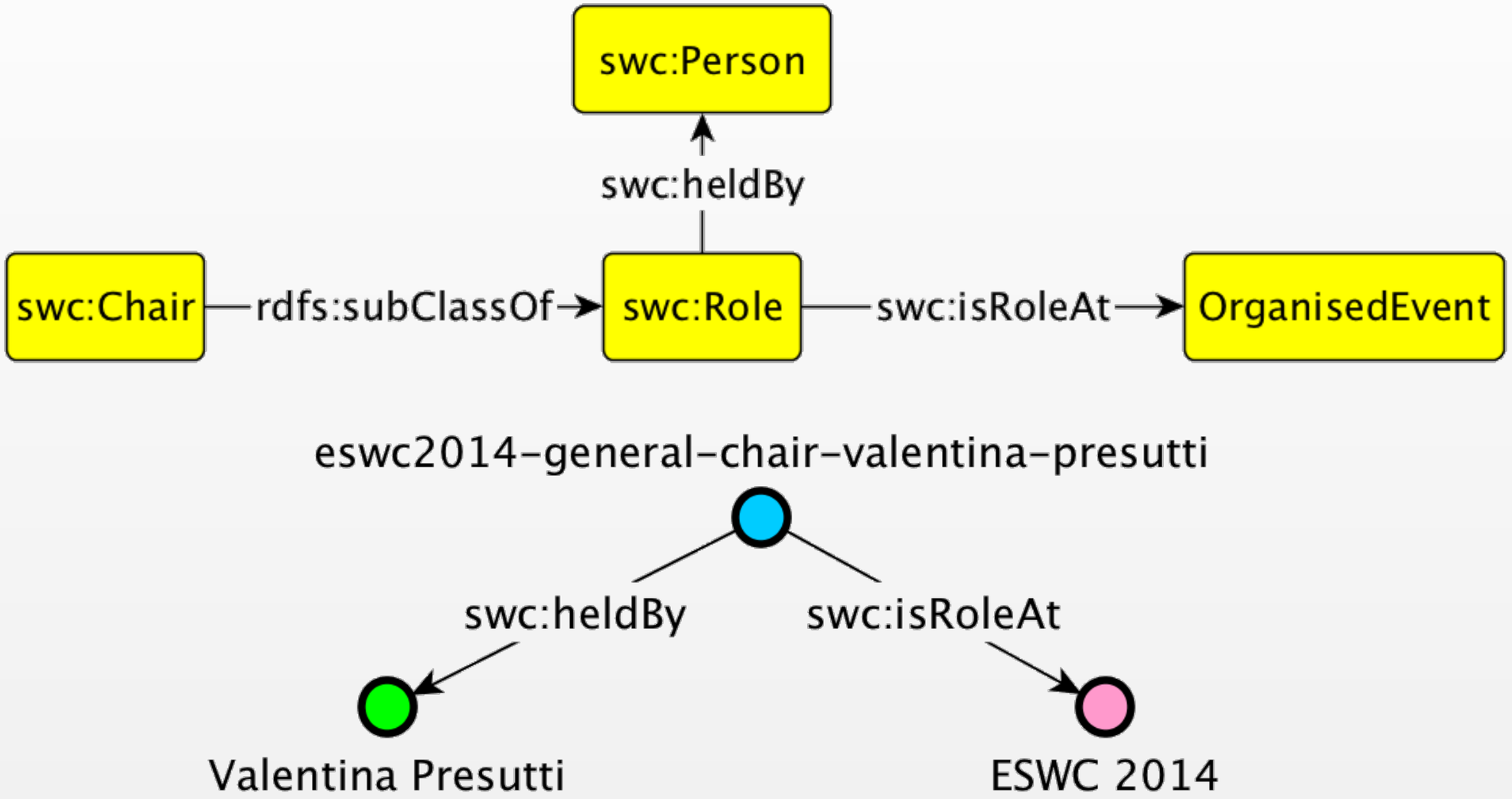
Time-indexed-situation ODP



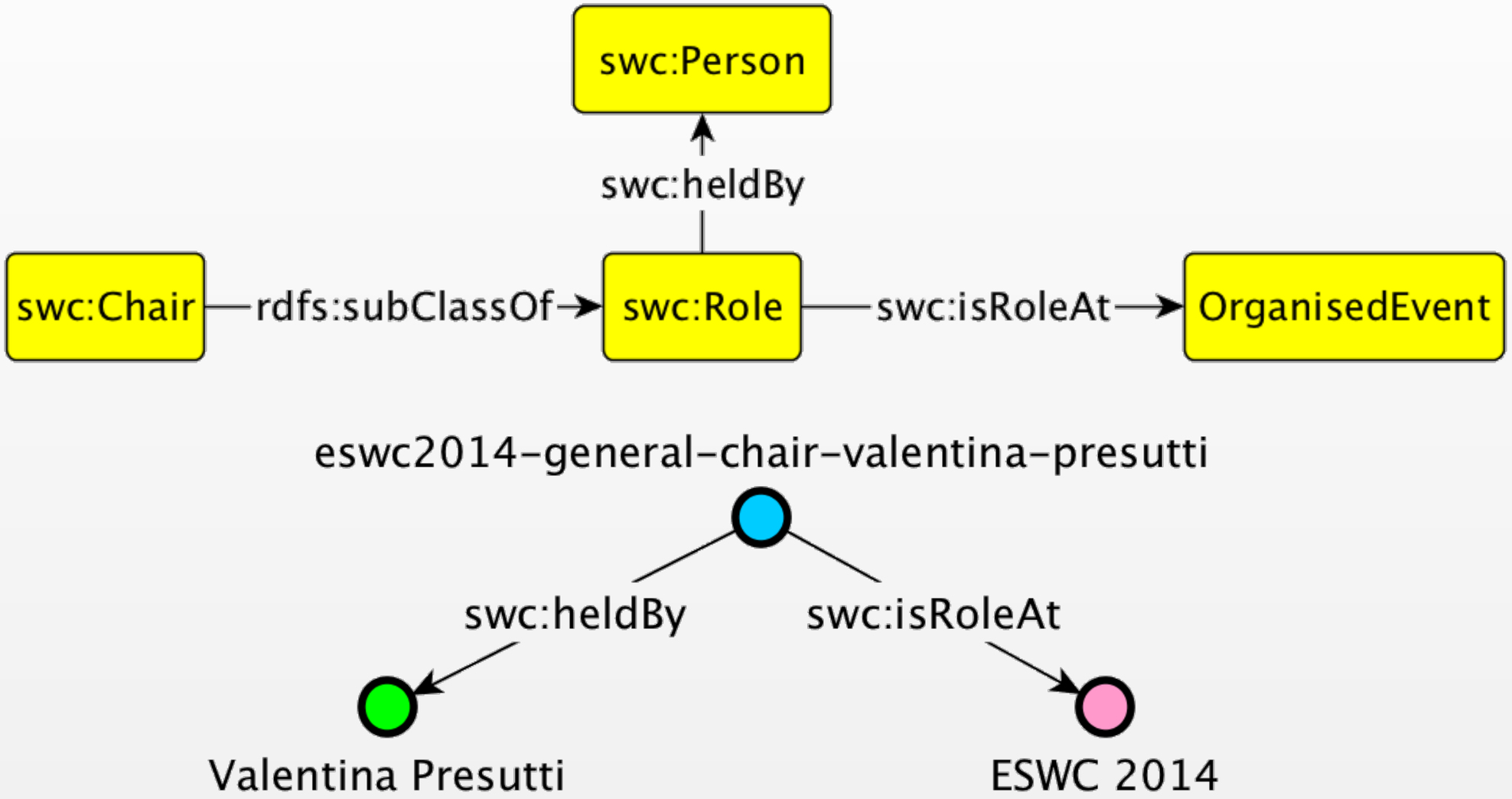
Roles: SWDF


swc:Person
swc:heldBy
swc:Chair
rdfs:subClassOf
swc:Role
swc:isRoleAt
OrganisedEvent

Roles: SWDF

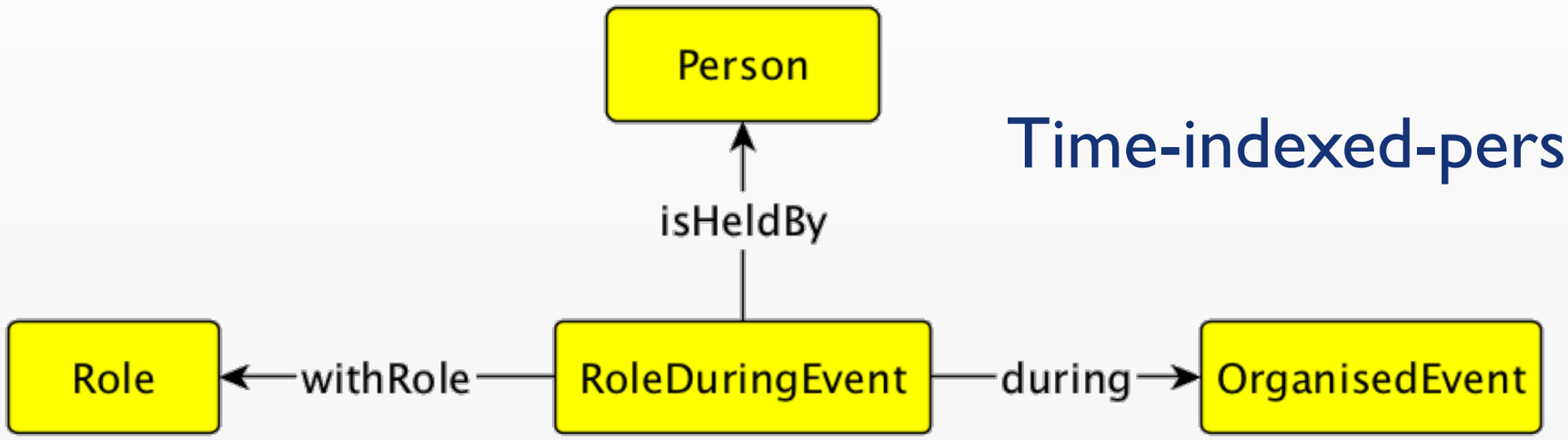

swc:Person
swc:heldBy
swc:Chair
rdfs:subClassOf
swc:Role
swc:isRoleAt
OrganisedEvent
eswc2014-general-chair-valentina-presutti
swc:heldBy
swc:isRoleAt
Valentina Presutti
ESWC 2014

Roles: SWDF

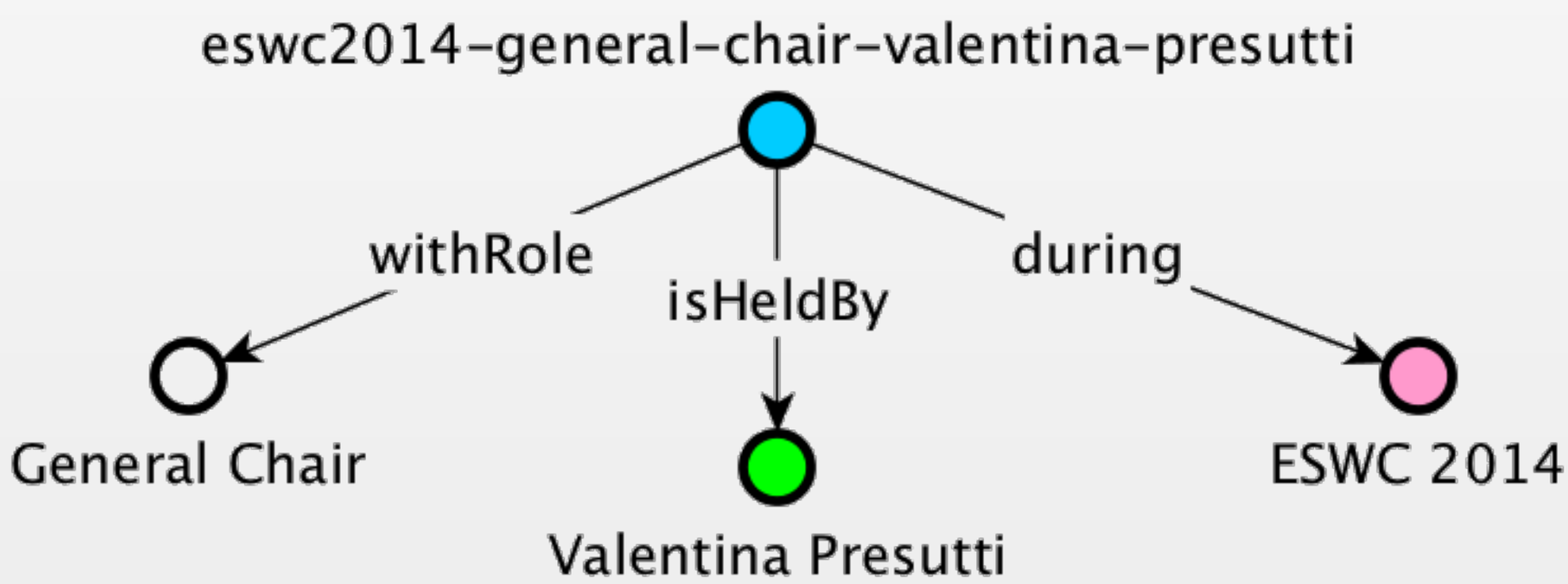


- Roles defined locally to each conference
  - 1,717 distinct individuals in the current dataset that truly represent a set of only 34 unique roles
- Difficult to answer queries where generalisation is needed
  - e.g. "Who were the general chairs at ESWC conferences?"

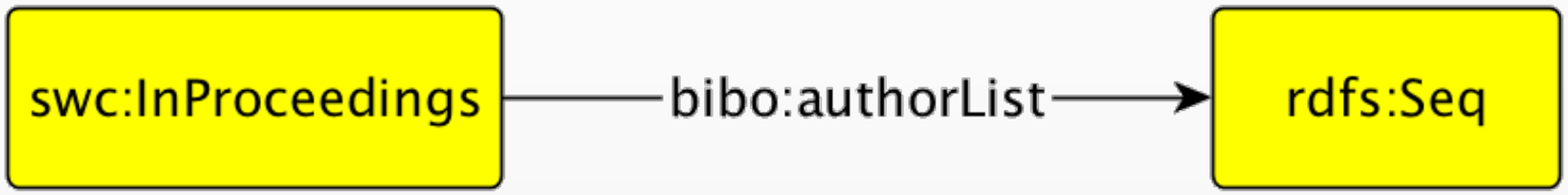
Roles: ScholarlyData



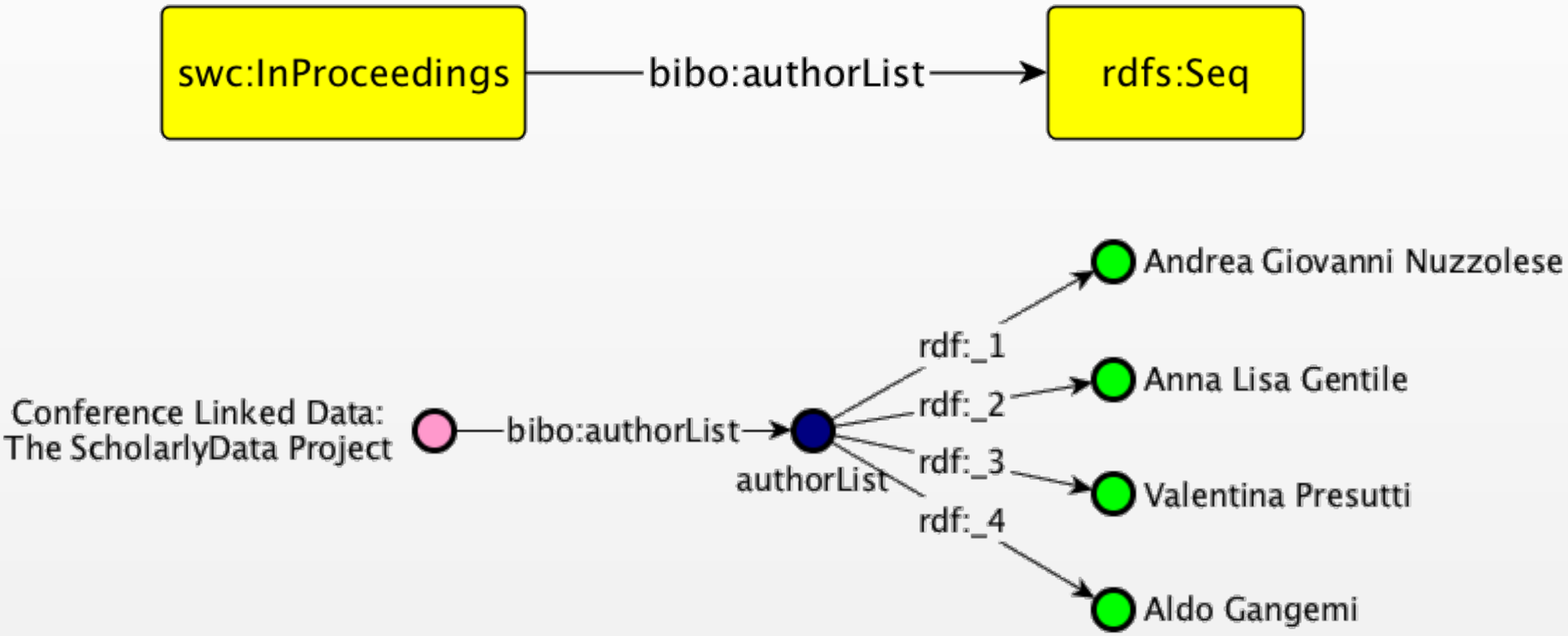
Time-indexed-person-role ODP



Lists: SWDF



Lists: SWDF

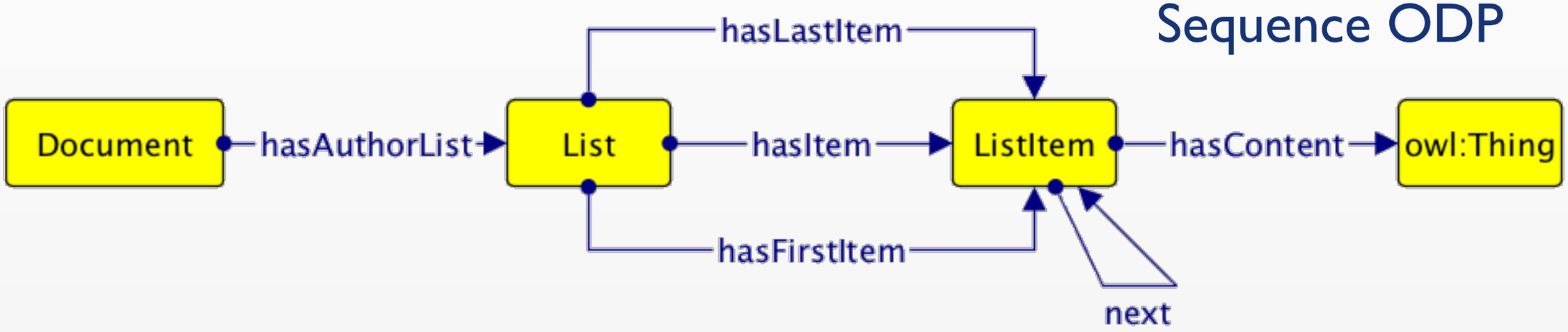


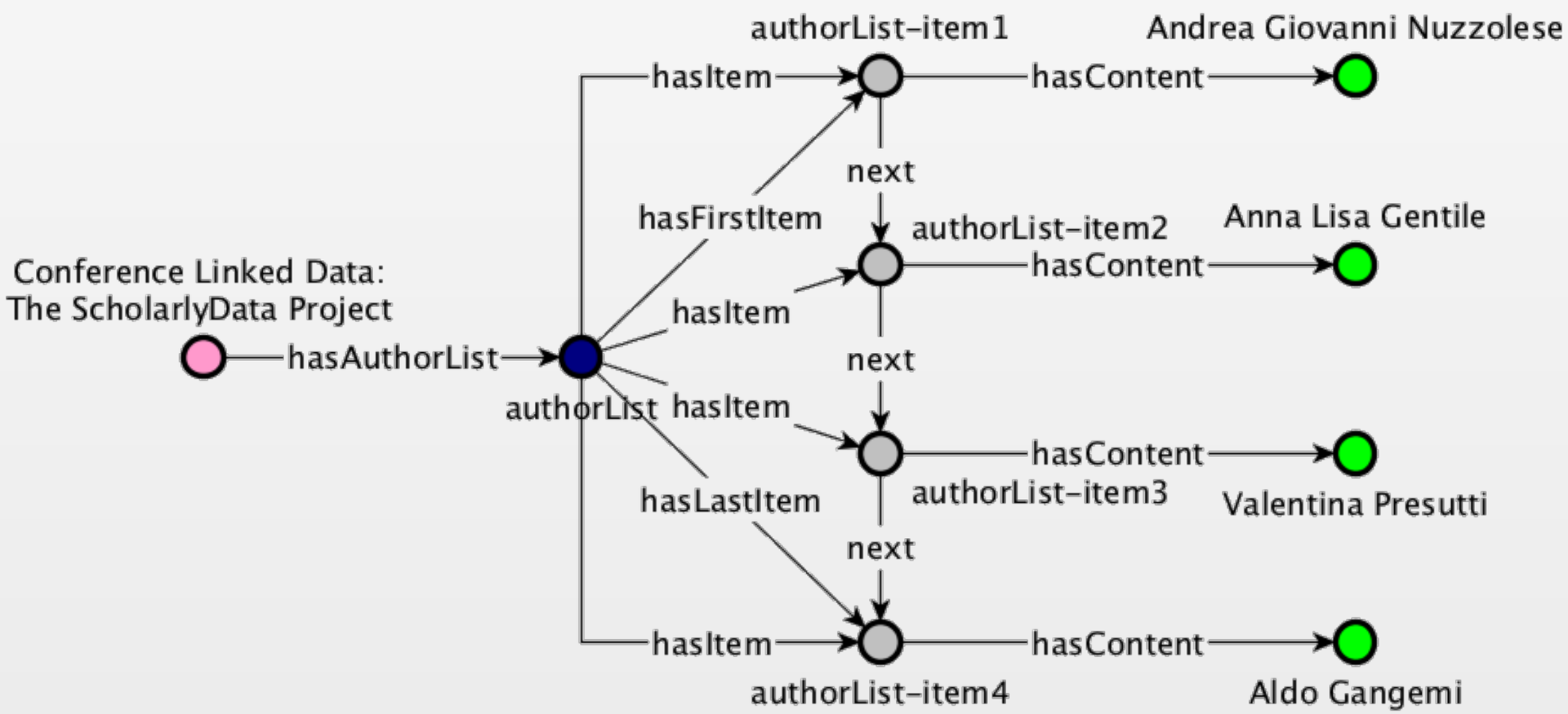
- OWL reasoning on RDF containers very hard [1]

[1] P. Ciccarese and S .Peroni, 2014. The Collections Ontology: creating and handling collections in OWL 2 DL frameworks. Semantic Web, 5(6), 515-529. DOI: 10.3233/SW-130121

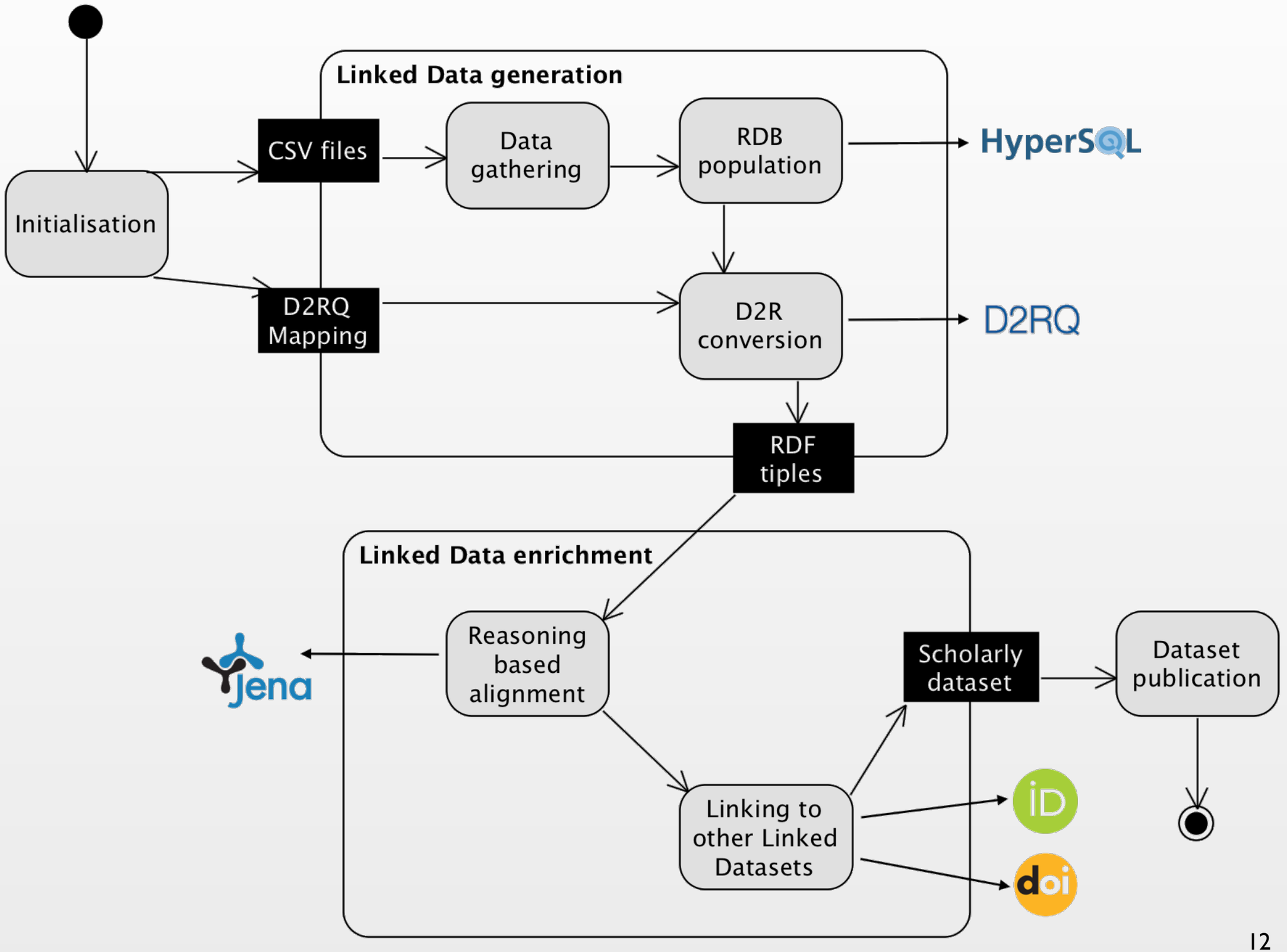
Lists: ScholarlyData

Sequence ODP



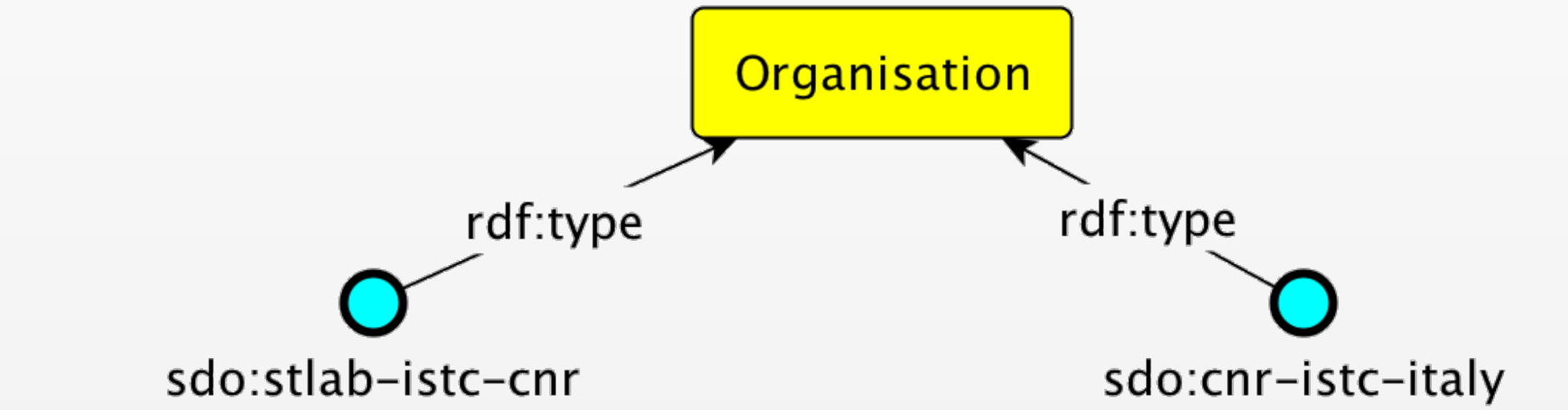


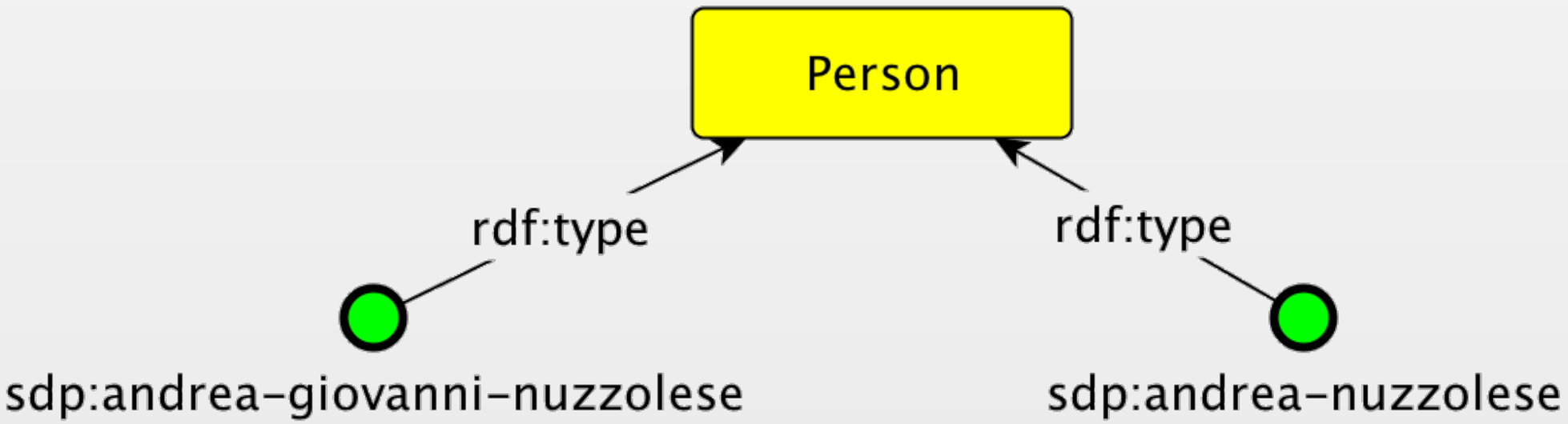
cLODg - conference Linked Open Data generator



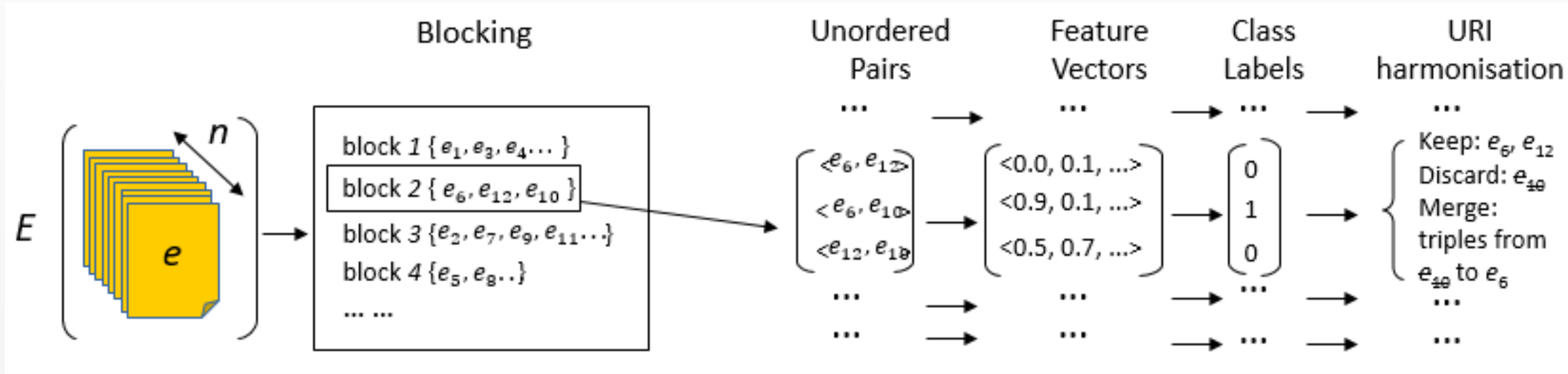
Entity deduplication

- Duplicates at instance level
- Most prominent entity types: PERSON, ORGANIZATION





Entity deduplication approach



- 0.86 precision for PERSON and 0.7 precision for ORGANIZATION

- Don't miss "Entity deduplication on ScholarlyData" at the main conference
  - Tuesday May 30 14:00 - Linked Data Session

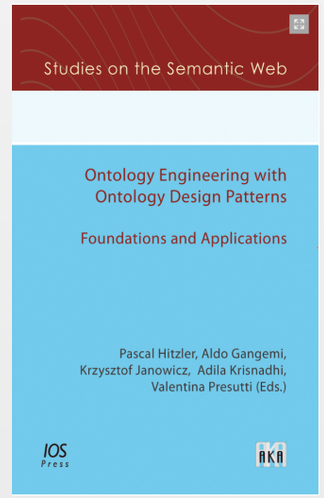
ScholarlyData

- Novel self-contained **conference-ontology**
  - fixes issues of SWC ontology
  - aligned to SWC, SPAR, W3C Organization Ontology, FOAF, SKOS
  - best ontology design practices [2]

- New dataset of scholarly facts
  - refactoring **of SWDF**

- **Novel** data generation workflow (cLODg)
  - support collaborative data generation
  - opensource on GitHub https://github.com/anuzzolese/cLODg2

https://w3id.org/scholarlydata/



[2] P. Hitzler, A. Gangemi, K. Janowicz, A. Krisnadhi, and V. Presutti, 2016. Ontology Engineering with Ontology Design Patterns: Foundations and Applications

Data details

- 1,128,618 triples about 93,519 individuals

Type	Individuals	Type	Individuals
`conf:TimeIndexedSituation`	20,998	`conf:RoleDuringEvent`	6,510
`conf:ListItem`	14,805	`conf:List`	4,463
`conf:AffiliationDuringEvent`	14,488	`conf:InProceedings`	4,393
`conf:Agent`	12,490	`conf:OrganisedEvent`	2,882
`conf:Person`	9,682	`conf:Organisation`	2,808

- 34 roles at global level instead of SWDF 1,717 roles at conference level
- instance level alignments to ORCID (~800 people) and DOI (~1K papers)

Resources

Data dumps

The RDF dumps of linked datasets

SPARQL endpoint

The endpoint where is possible to query data via SPARQL

Conference Ontology

A novel data model which improves the Semantic Web Conference Ontology, adopting best ontology design practices

Datahub

ScholarlyData.org datasets on datahub.io

cLODg code

The source code of cLODg on GitHub. cLODg is a software that generates conference Linked Open Data generator from EasyChair dumps

Dataset loader

Service for uploading new scholarly datasets and publish them as Linked Oped Data on ScholarlyData.org

Future work

- Evaluation of the resource
- Introduction of more sophisticated components to deal with instance matching in the cLODg workflow
- More linking, e.g. OpenCitations
- Fostering collaboration with Conference Management System providers to provide cLODg as a build-in facility in the systems

Thank you