

Ontology-based Generation of Personalised Data Management Systems: an Application to Experimental Particle Physics

PhD thesis defense of Blerina GKOTSE

Blerina.Gkotse@cern.ch

MINES ParisTech, PSL University, France

25 September 2020

Jury:

Laura GONELLA, Rapporteur

Jean-Baptiste LAMY, Rapporteur

Laurent DUSSEAU, Examineur

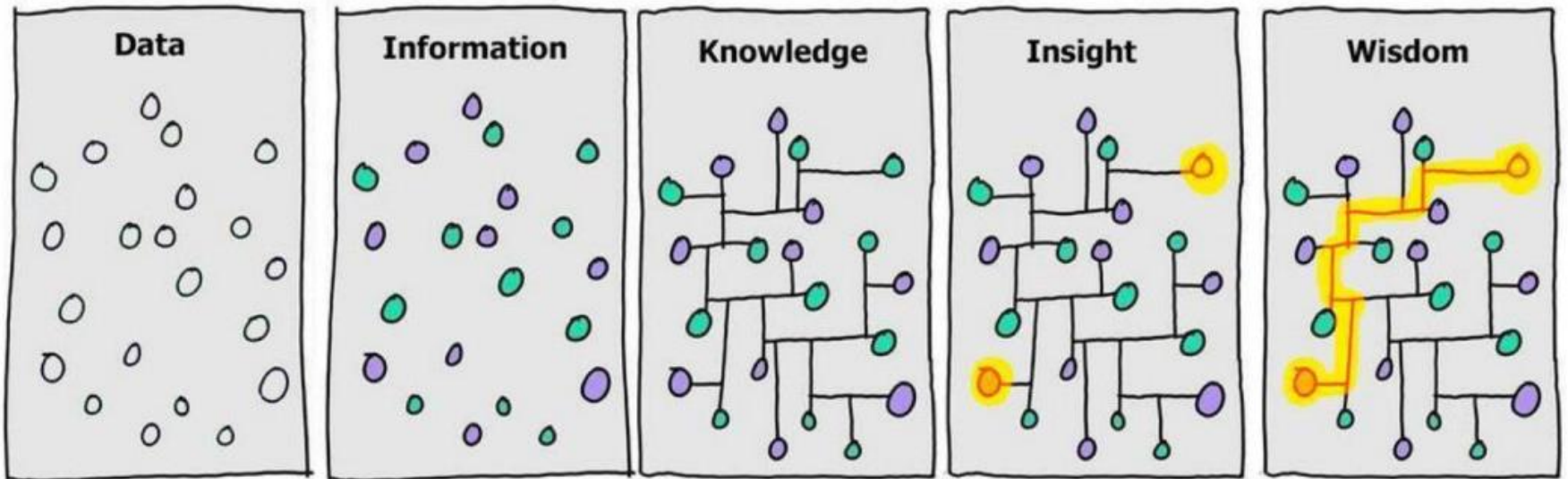
Theodora VARVARIGOU, Examineur

Pierre JOUVELOT, Directeur de thèse

Federico RAVOTTI, Maître de thèse



Motivation



©gapingvoid

From data to wisdom

Context

Experimental Particle Physics (EPP) Web Semantics



Large Hadron Collider (LHC)



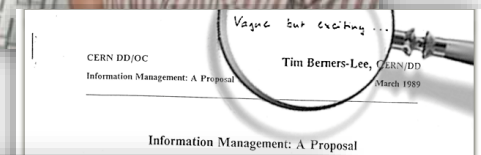
Tracker



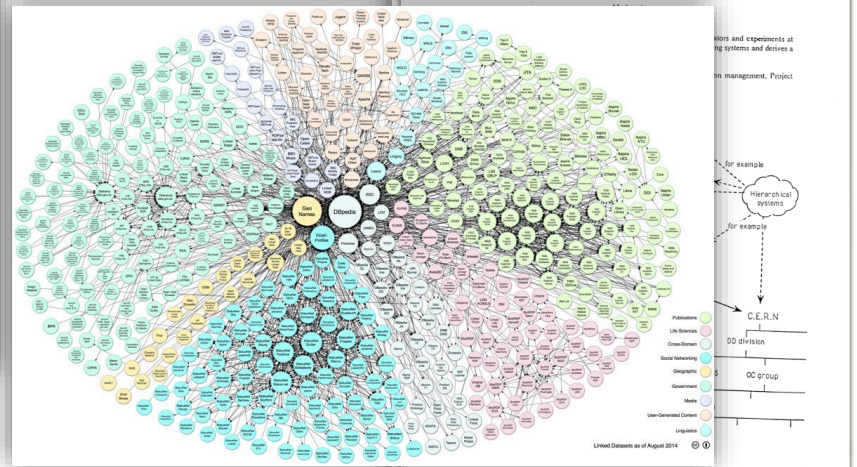
Compact Muon Solenoid (CMS)



Tim Berners Lee



Information Management: A Proposal



Linked data

First Proposal for WWW

Web Semantics

Computers don't understand semantic meaning

E.g., *“My mouse is broken...”*



Ontology Origin

Deriving from ancient Greek:

Ontology = On (ὄν) + logos (λόγος)

In philosophy:

subject of existence, science of being,
"what exists" in the world

In computer science:

formal description and classification of "what exists",
what can be represented as a knowledge fact

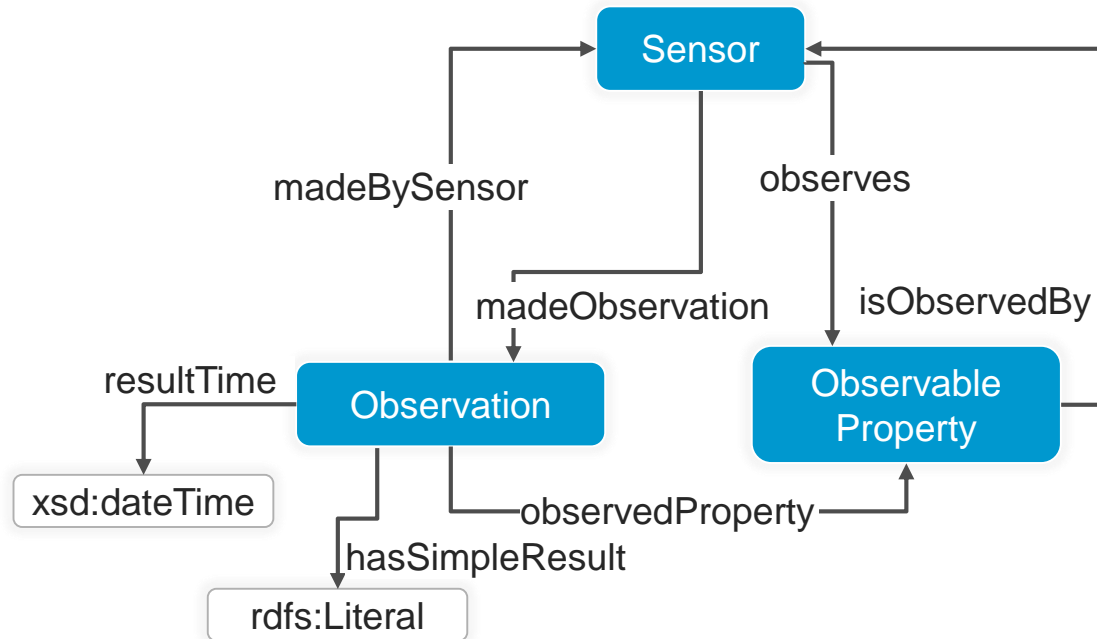


©shutterstock.com

Aristotle

Ontology: Structure

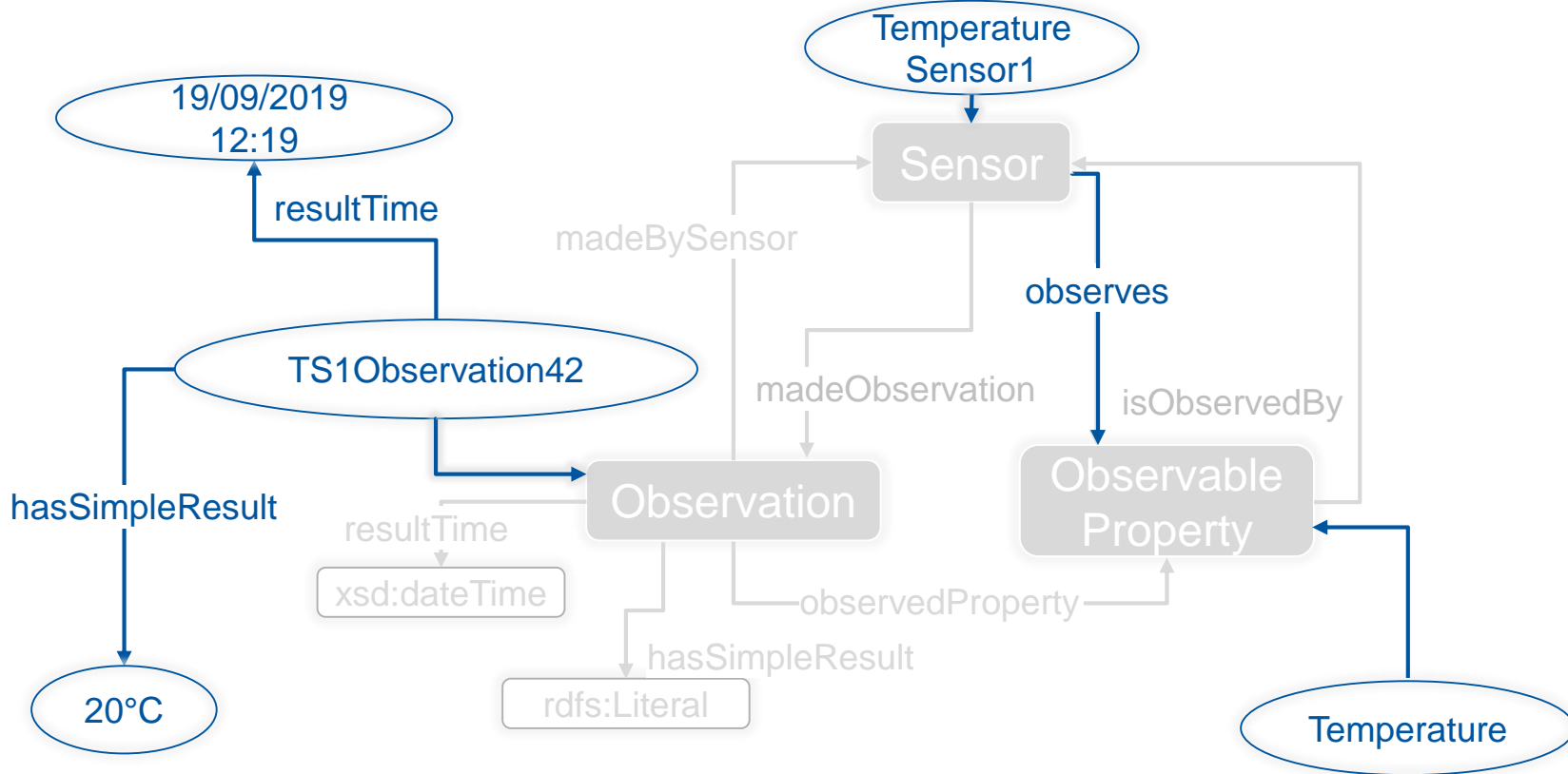
- **Class:** a set of entities of a specific domain of knowledge
- **Relation:** a semantic link among classes, also called **object property**
- **Property:** attribute of specific type, also called **data property**



Excerpt from SOSA (Sensors, Observations, Samples and Actuators) ontology

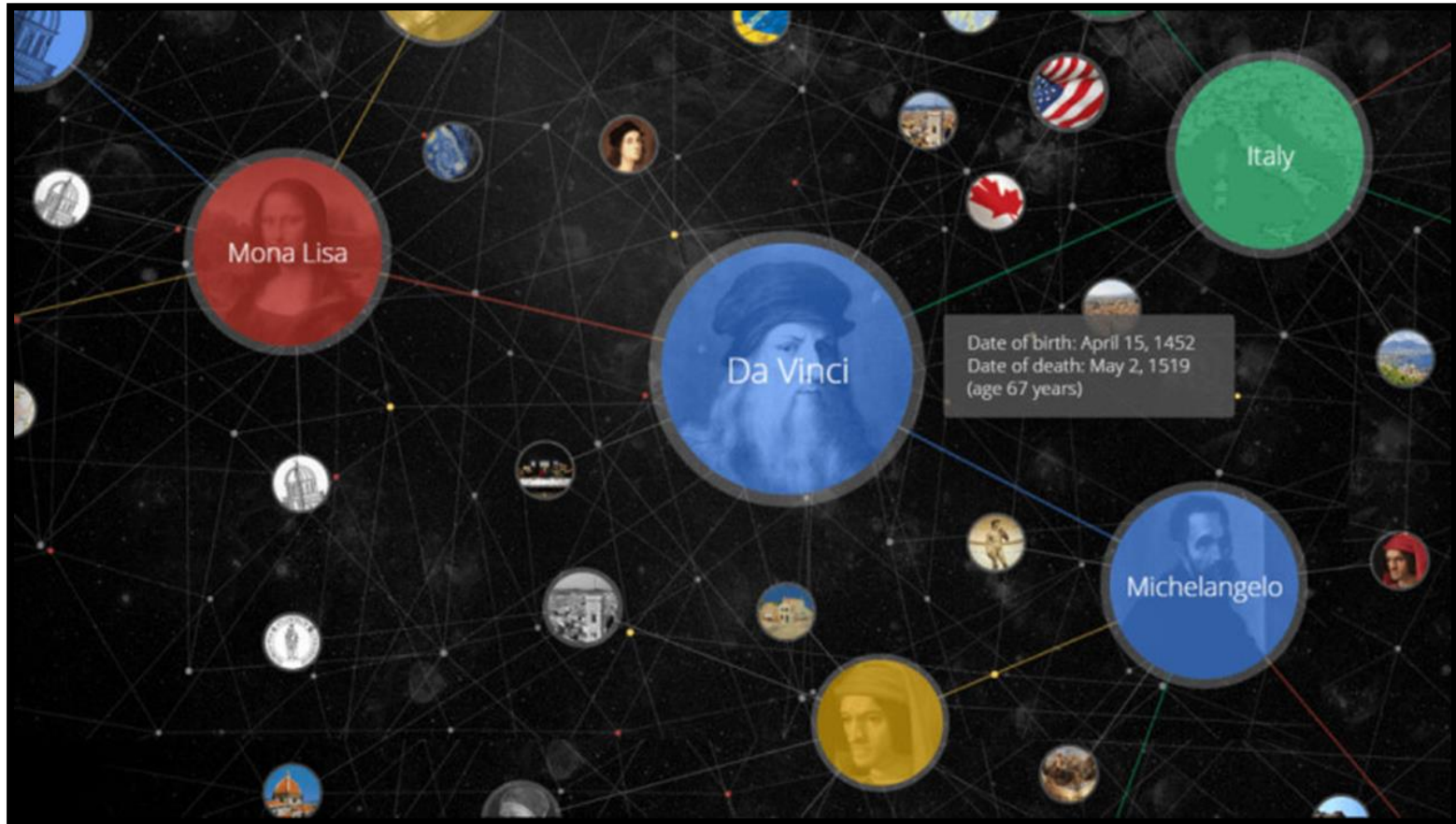
Knowledge Base (KB)

- Individual instances of domain's classes
- RDF triples (Resource Description Framework)
- Stored in triple stores



Knowledge Graph (KG)

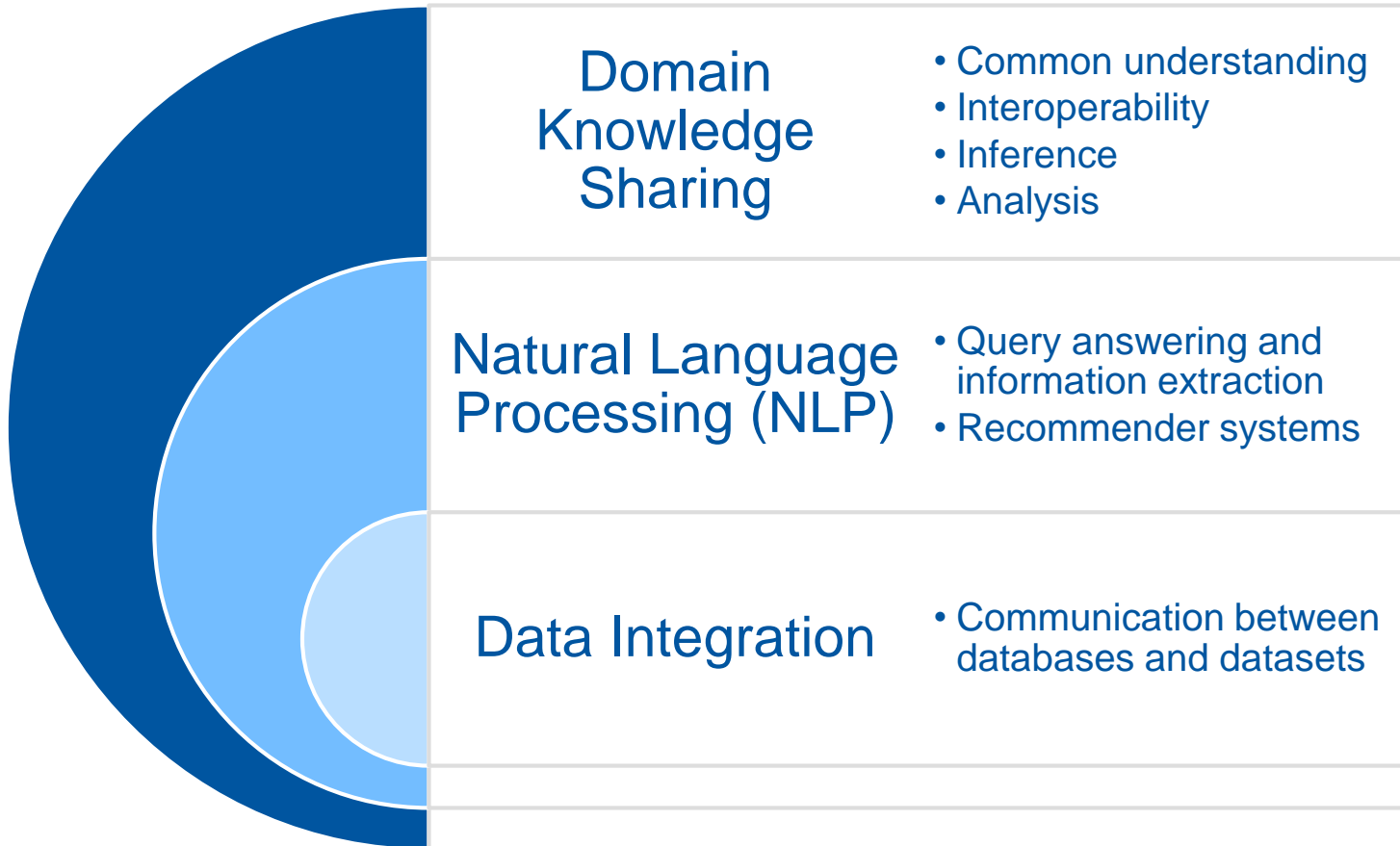
- Term coined by Google (2012)
- Knowledge bases from a variety of sources
- Used in academia and industry



Google Knowledge Graph

©searchengineland.com

Ontology Goals

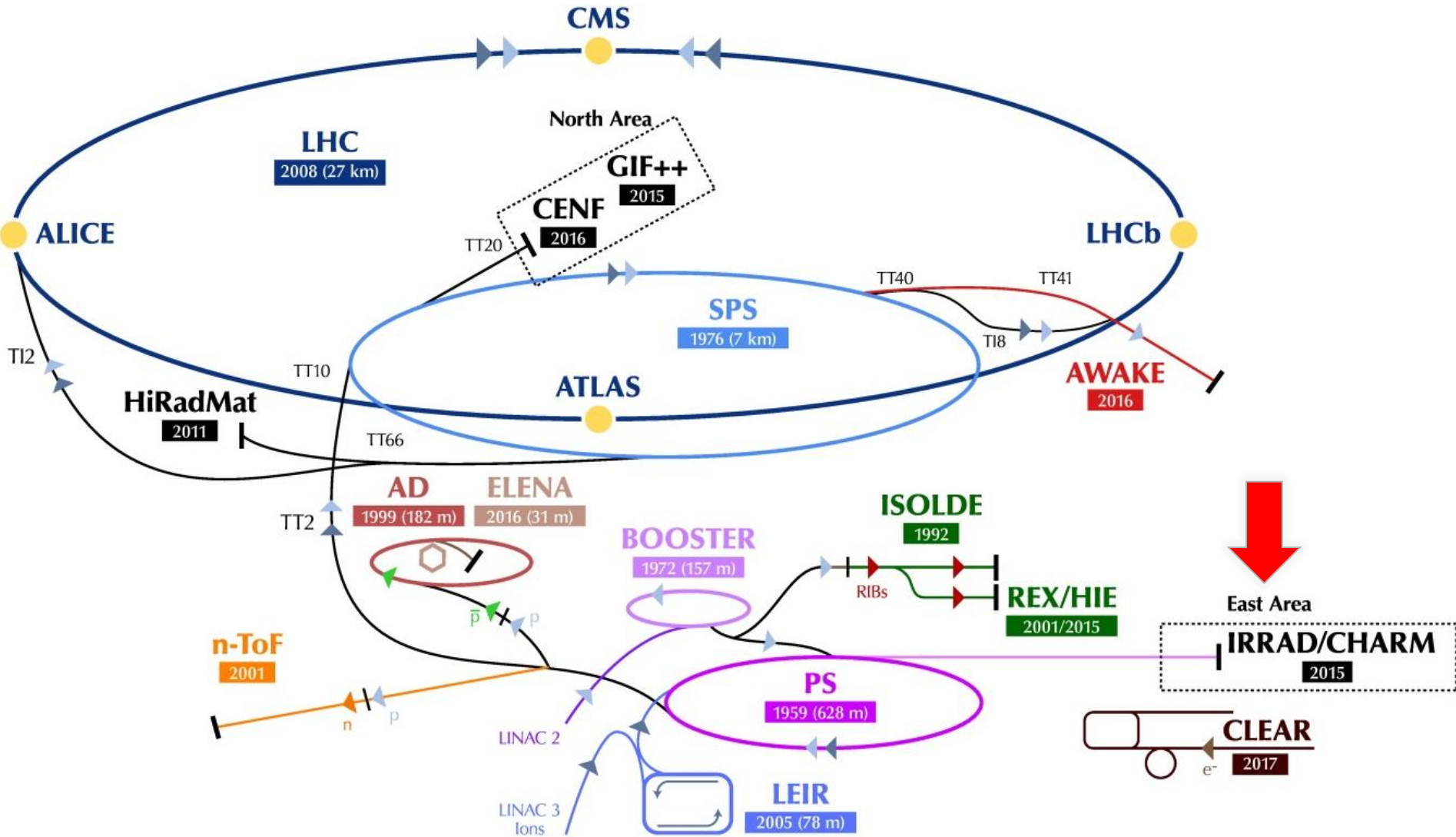


European Laboratory of Particle Physics (CERN)



LHC (Large Hadron Collider)

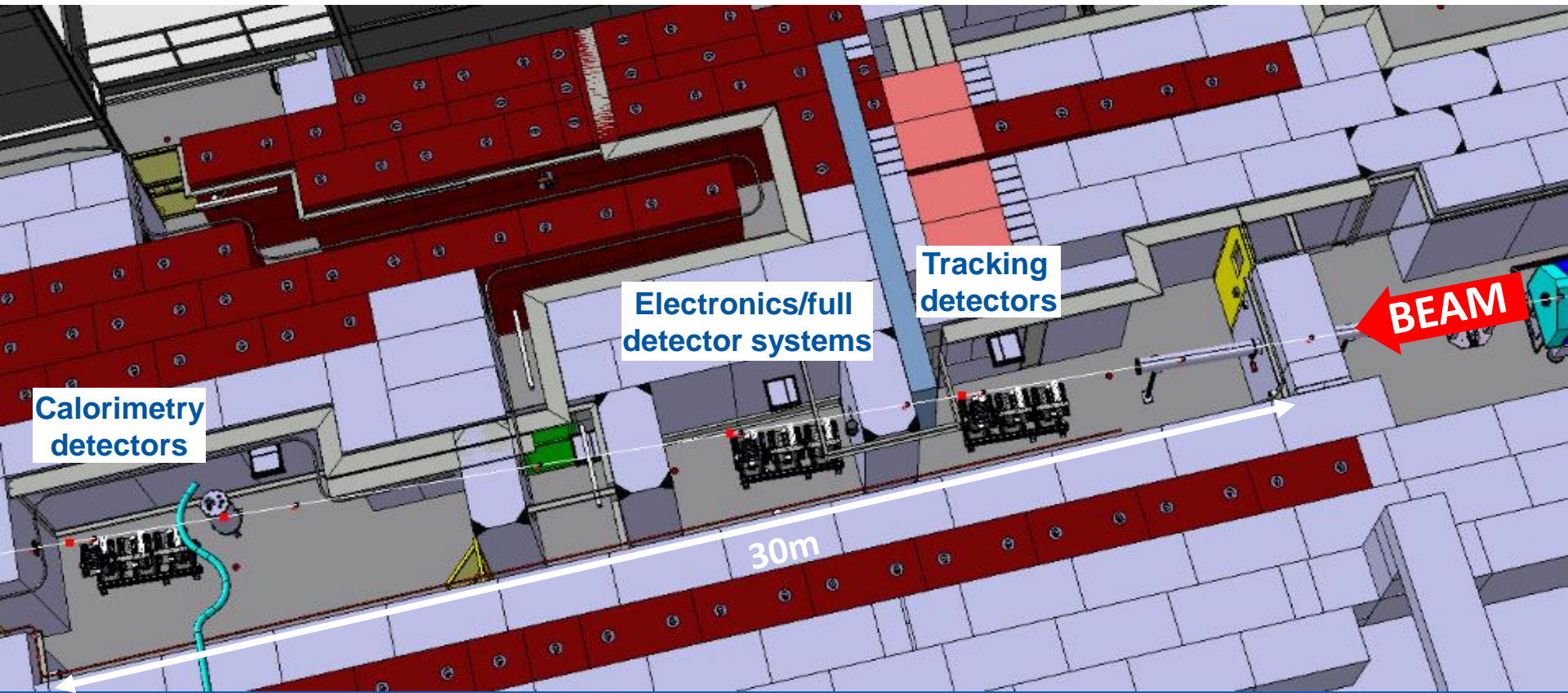
CERN Accelerators Complex



©cds.cern.ch

CERN Proton Irradiation Facility (IRRAD)

- Reference facility for proton irradiation experiments
- Testing of components of EPP experiments



IRRAD Experiments

In 2018: **81** irradiation experiments,
97 users, **792** samples, **405** dosimeters,
2056 dosimetry measurements

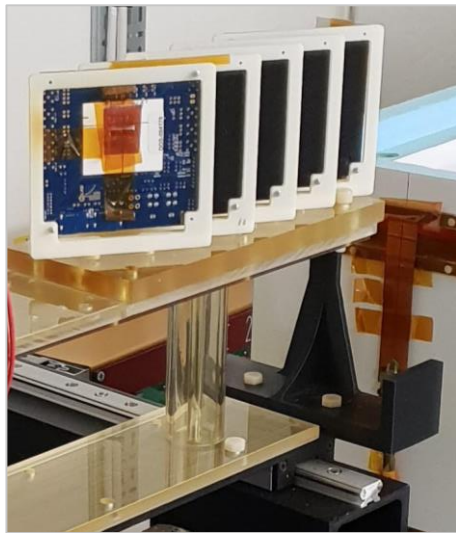
Proper data
management necessary



Piezo actuators



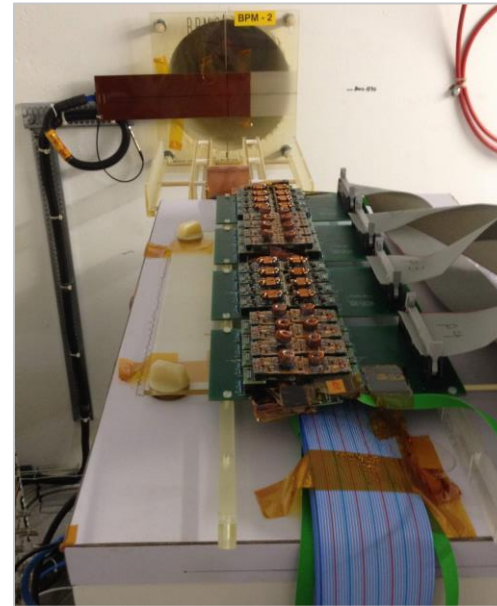
ECAL crystal



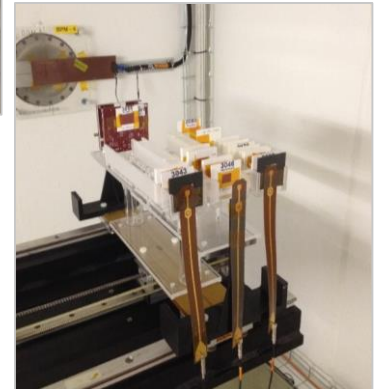
RD53A modules

©ps-irrad.cern.ch

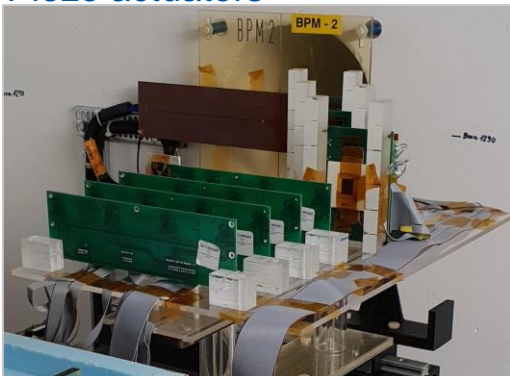
DC/DC converters



Full-tracking detector module



Si pad diode



CLARO ASIC

IRRAD Data Manager (IDM)



IRRAD Data Manager

Irradiation Experiments

ID	Irradiation title	Availability	No. registered irradiated samples	Radiation/No.CAL Length Consistency (%)	No. Beams	Responsible person	Status	Actions
81	FCC-RADIATION [2]	01/04/2018	4/8	4.189 / 3.893	0	Georg Garbe	Completed	Validated Deleted Archived Refresh
841	Photo-diode irradiation [2]	29/06/2018	3/3	9.111 / 4.212	0	Lauri Järvelä-Ontaranta	Completed	Validated Deleted Archived Refresh
1483	Heavy Ion Irradiation for silicon defect spectroscopy [42]	14/11/2018	4/4	0.232 / 0.088	0	Ilkka Mäkelä	Validated	Validated Deleted Archived Refresh
1482	NEEL Calibration - Heavy ION 2018 - RADMON [42]	14/11/2018	1/1	0.085 / 0.063	0	Glennys Peticola	Validated	Validated Deleted Archived Refresh
1481	NEEL Calibration - Heavy ION 2018 [42]	14/11/2018	5/5	3.228 / 0.995	0	Glennys Peticola	Validated	Validated Deleted Archived Refresh

IRRAD Data Manager

Dosimetry results for SET-003252 (ULTEM1000)

Dosimeter	Dimensions (mm ³)	Date In	Date Out	SEC	Accumulated fluence	Error(%)	Comments
DOS-004003	50x30	18/04/2018 20:02	05/09/2018 03:00	1.45e+10	9.79e+16	7	
DOS-004151	50x30	12/09/2018 13:25		0.00e+00		None	

Dosimeter dimensions (mm³)
50x30mm³

Total accumulated fluence
9.790e+16 Protons/cm²

IRRAD Data Manager

TREC Data of Sample SET-003122

Characteristics

- Length: 1 cm
- Width: 0.1 cm
- Height: 0.002 cm
- Weight: 0.114g
- Family: 1
- Material: 1

Registration

Planning

Follow-up

Dosimetry Results

Traceability

History

Sample details

Sample type details

Sample dimensions

Name	Length (mm)	Weight (g)	Beam	Beam
L1	0.03	0.001	AT15	AT15
L2	0.03	0.001	AT15	AT15
L3	0.03	0.001	AT15	AT15

IRRAD Data Manager

Irradiation Status

Updated at	Sample	Dosimeter	Date In - Date Out	IRRAD table	Table position	Accumulated fluence	SEC	Updated by	Status	In Beam	Actions
15/11/2018	SET-003899	DOS-004211	15/11/2018 16:11	IRRAD19	Center		272851	irradiation.facilities@cern.ch	Registered	<input type="checkbox"/>	Validated Deleted Archived Refresh
15/11/2018	SET-003900	DOS-004211	15/11/2018 16:11	IRRAD19	Center		272851	irradiation.facilities@cern.ch	Registered	<input type="checkbox"/>	Validated Deleted Archived Refresh
15/11/2018	SET-003901	DOS-004211	15/11/2018 16:11	IRRAD19	Center		272851	irradiation.facilities@cern.ch	Registered	<input type="checkbox"/>	Validated Deleted Archived Refresh
15/11/2018	SET-003902	DOS-004211	15/11/2018 16:11	IRRAD19	Center		272851	irradiation.facilities@cern.ch	Registered	<input type="checkbox"/>	Validated Deleted Archived Refresh
15/11/2018	SET-003903	DOS-004211	15/11/2018 16:11	IRRAD19	Center		272851	irradiation.facilities@cern.ch	Registered	<input type="checkbox"/>	Validated Deleted Archived Refresh

IRRAD Data Manager

3D pixel for ATLAS ITK

Experiment Details

Title: 3D pixel for ATLAS ITK

Description: Study of radiation hardness of 3D silicon pixel sensors for the innermost pixel layer of ATLAS ITK with fluences up to 2e16 protons/cm². Both FEEM prototypes and modules with the new REDCA readout chip are going to be tested.

Responsible person: jern.samp@cern.ch

Installation type: Proton

Sample Details

Category: Positive Custom

Type: Beam temperature

Installation area: 20x20mm²

Material: Silicon sensor + readout chip bump bonded

Fluence: 1e16 Protons/cm² 2e16 Protons/cm² 3e16 Protons/cm²

Number of samples: 1

Additional comments: Important to characterize the new REDCA readout modules for ATLAS ITK pixel before the PS/PSB shutdown in 2019/20.



Irradiation Experiments



CERN Large Hadron Collider (LHC)
©cds.cern.ch



NASA's Orion spacecraft
©arstechnica.com



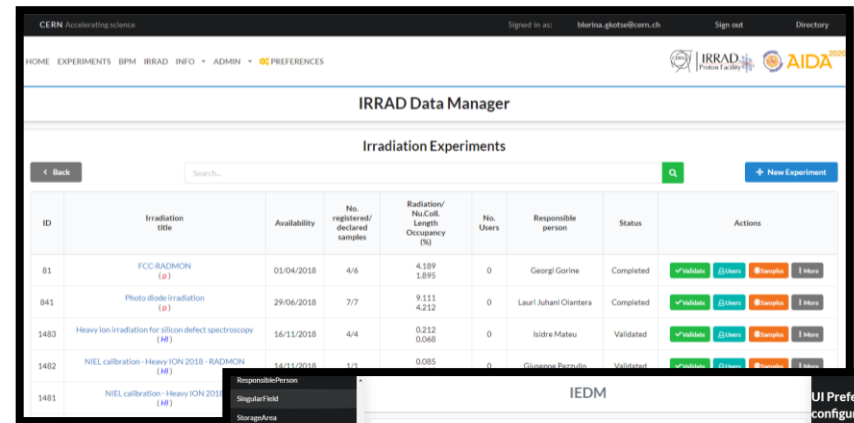
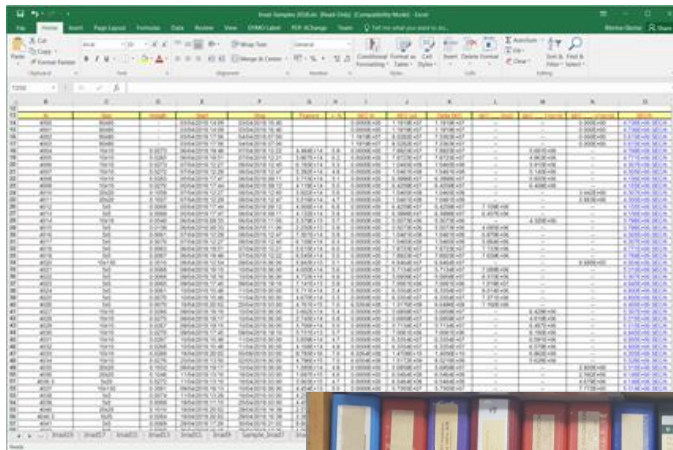
Bottle sterilization by electron beam irradiation
©Courtesy Frank-Holm Roegner



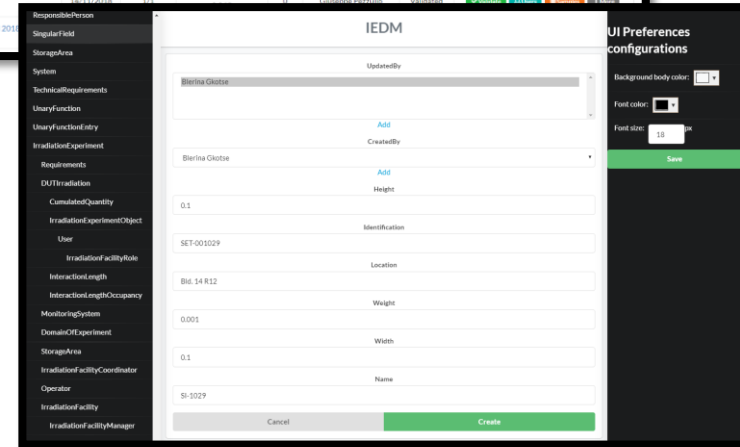
Clinac® iX System linear accelerator
©medicalexp.com

Bridging the Gap in Data Management

- Knowledge sharing among communities
 - Small experiments, no strong IT support
 - User Experience
- ✓ Web Semantics
 - ✓ Automatic generation of web applications
 - ✓ UI personalization



ID	Irradiation title	Availability	No. registered/ declared samples	Radiation/ No.Coll. Length Occupancy (%)	No. Users	Responsible person	Status	Actions
81	FCC-BADMON (p)	01/04/2018	4/6	4.189 1.895	0	Georgi Gorine	Completed	✓ Addable ✓ Usable ✓ Searchable ✓ Hides
841	Photo diode Irradiation (p)	29/06/2018	7/7	9.111 4.212	0	Lauri Ahanli Olantera	Completed	✓ Addable ✓ Usable ✓ Searchable ✓ Hides
1483	Heavy Ion Irradiation for silicon defect spectroscopy (H)	16/11/2018	4/4	0.212 0.068	0	Ildre Mateu	Validated	✓ Addable ✓ Usable ✓ Searchable ✓ Hides
1482	NIEL calibration - Heavy ION 2018 - BADMON (H)	14/11/2018	1/1	0.085	0	Christina Doro	Validated	✓ Addable ✓ Usable ✓ Searchable ✓ Hides
1481	NIEL calibration - Heavy ION 2018 - BADMON (H)							✓ Addable ✓ Usable ✓ Searchable ✓ Hides



IEDM

UpdateBy:

CreateBy:

Height:

Identification:

Location:

Weight:

Width:

Name:

Cancel Create

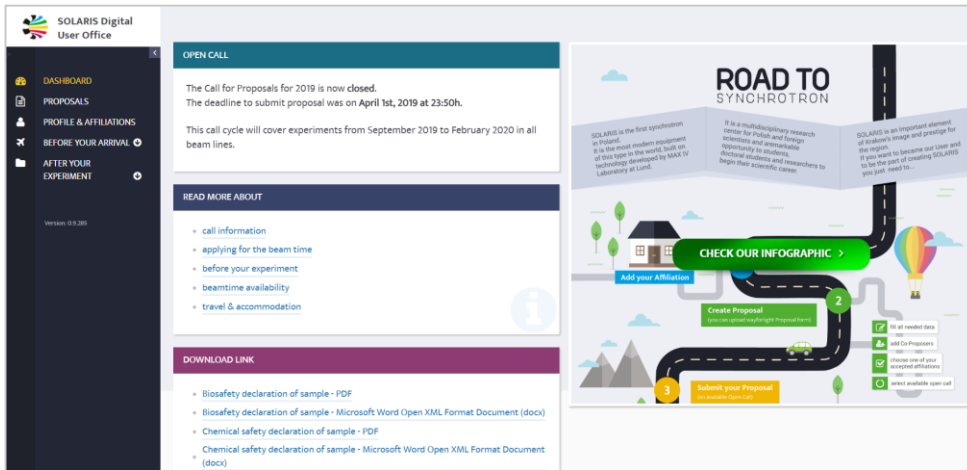
Outline

- **Context**
 - **Web Semantics**
 - **Experimental Particle Physics**
- **Irradiation Experiment Data Management Ontology (IEDM)**
- **Automatic Generation of Web Applications from Ontologies**
- **UI Personalization with Ontology Embeddings**

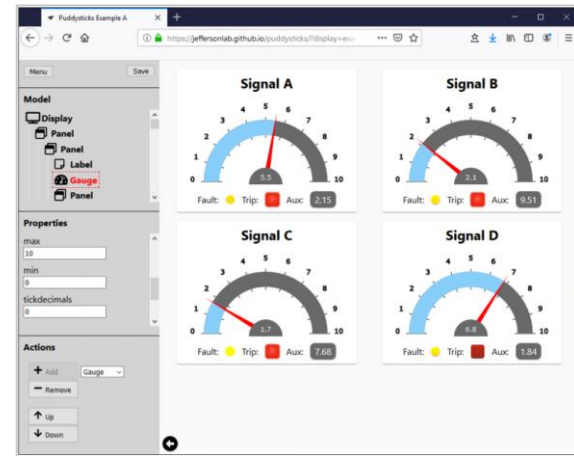
Data Management in EPP

Available tools for:

- Information registration
- Infrastructure availability
- Monitoring and visualization
- Reporting



SOLARIS Digital User Office



Web Extensible Display Manager used at Jefferson Lab for control-systems display

- **No fully integrated solution for the follow-up of experiments' life cycle**
- **No use of web semantics**

Domain Ontologies in EPP

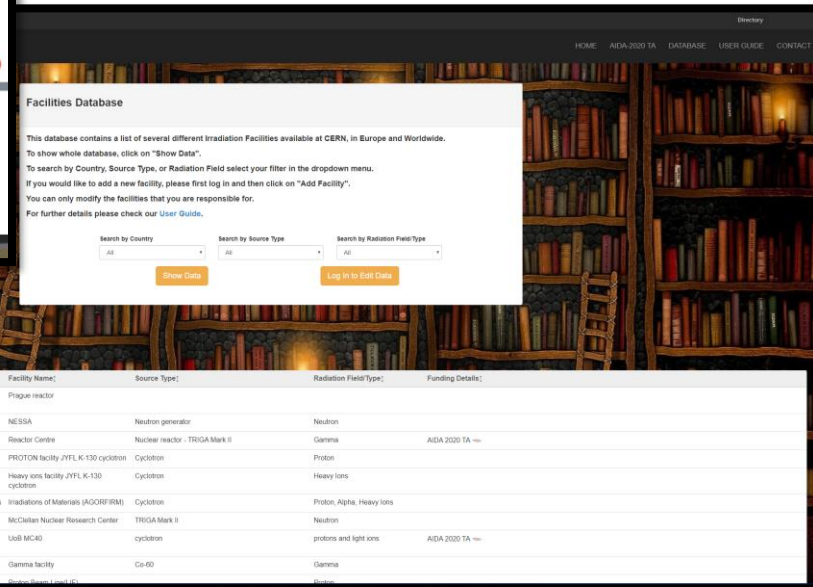
Ontology	Authors	Content	Domain
Web Physics Ontology	V. Vjetkovic	physics equations and relations	Physics
Ontology Design Pattern	D. Carral, <i>et al.</i>	analysis of particle physics data	Physics
Synchrotron Ontology	J. Szota-Pachowic	synchrotron components	Physics
Radiation Protection Ontology	C. Barki, <i>et al.</i>	irradiation experiments focusing on medicine	Medicine
EXPO – Ontology of Scientific Experiments	L. Soldatova, <i>et al.</i>	scientific experiments	General

No ontology for the data management of irradiation experiments

Irradiation Experiments' Data Research

- Research and compilation of irradiation experiments' and related facilities' data
- Domain experts interviewed (~10)

Online DB platform developed containing the collected data (217 facilities up to date)
www.cern.ch/irradiation-facilities

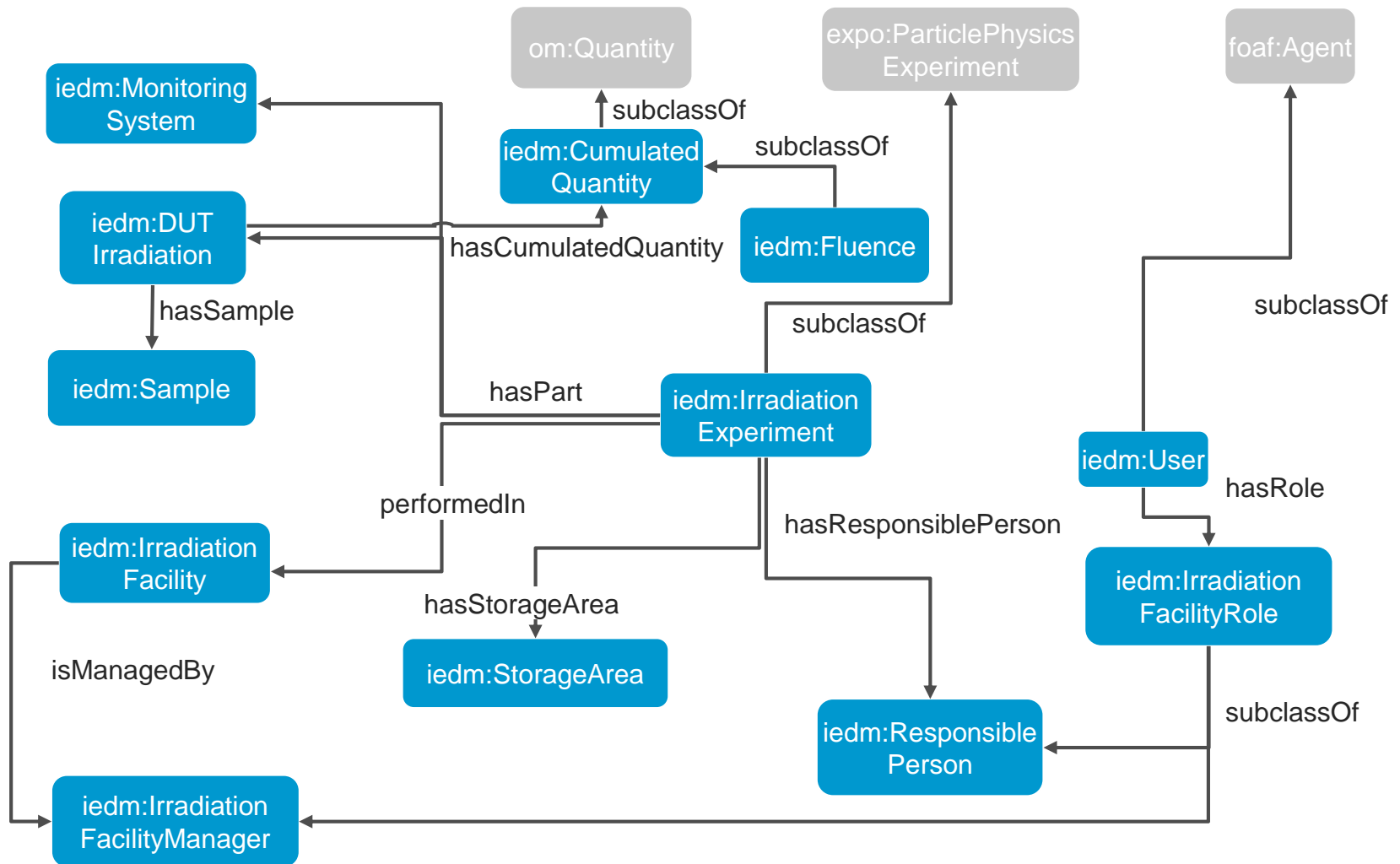


Irradiation Experiment Data Management (IEDM)

Three foundational ontologies

Name	Authors	Concepts
EXPO – Ontology of Scientific Experiments	L. Soldatova, <i>et al.</i>	scientific experiments
OM-2 – Units of Measure ontology	H. Rijgersberg, <i>et al.</i>	experimental quantities and physical units
FOAF – Friend Of A Friend Ontology	D. Brickley, <i>et al.</i>	networks of people, activities and their relations

IEDM: Core Classes



IEDM Online Documentation

<http://cern.ch/iedm>



Irradiation Experiment Data Management Ontology (IEDM)

language [en](#)

Release 2019-03-10

This version:

<https://gitlab.cern.ch/bgkotse/iedm/raw/master/iedm.owl>

Authors:

B. Gkotse, P. Jouvelot and F. Ravotti

Download serialization:

[Format: JSON LD](#) [Format: RDF/XML](#) [Format: N Triples](#) [Format: TTL](#)

License:

[License](#) [license name goes here](#)

Visualization:

[Visualize with WebVowl](#)

Cite as:

B. Gkotse, P. Jouvelot and F. Ravotti. Irradiation Experiment Data Management Ontology (IEDM). Revision: 0.1. <https://gitlab.cern.ch/bgkotse/iedm/raw/master/iedm.owl>

Abstract

Irradiation experiments (IE) are an essential step in the development of High-Energy Physics particle accelerators and detectors. They are used to assess the radiation hardness of experimental devices by simulating, in a short time, the common long-term degradation effects due to high-energy particles. Usually carried out with ionizing radiation, these complex processes require highly specialized infrastructures called "irradiation facilities". Aiming to promote knowledge sharing and digital management of IEs, we introduce IEDM, a new Irradiation Experiment Data Management ontology.

Table of contents

- 1. [Introduction](#)
 - 1.1. [Namespace declarations](#)
- 2. [Irradiation Experiment Data Management Ontology: Overview](#)
- 3. [Irradiation Experiment Data Management Ontology: Description](#)
- 4. [Cross reference for Irradiation Experiment Data Management Ontology classes, properties and dataproperties](#)
 - 4.1. [Classes](#)
 - 4.2. [Object Properties](#)
 - 4.3. [Data Properties](#)
 - 4.4. [Named Individuals](#)
- 5. [References](#)
- 6. [Acknowledgements](#)

115 classes
941 annotations
24 object properties
16 data properties



Documentation generated using
WIDOCO.
D. Garijo, "Widoco: a wizard for
documenting ontologies" ISWC 2017

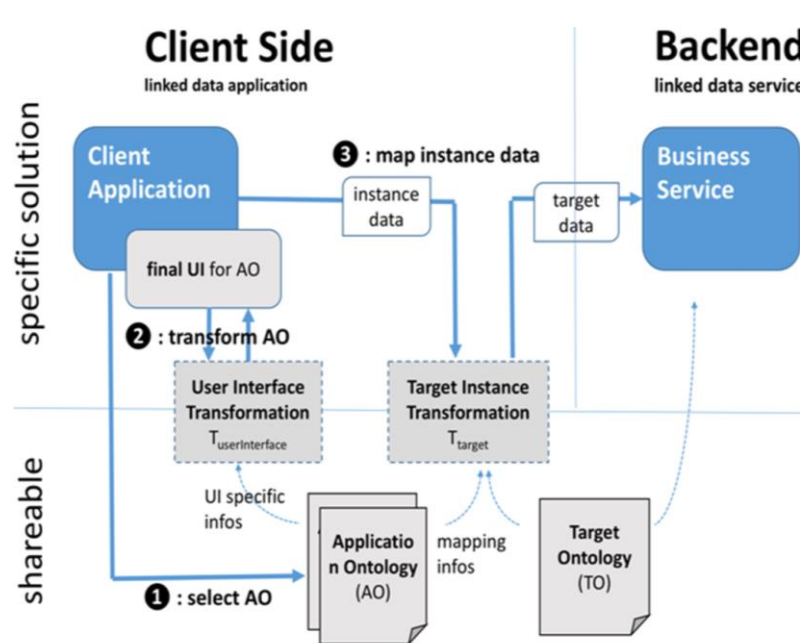
Outline

- Web Semantics
- Experimental Particle Physics
- Irradiation Experiment Data Management Ontology (IEDM)
- **Automatic Generation of Web Applications from Ontologies**
- UI Personalization with Ontology Embeddings

Ontology-based User Interface (UI) Generation

Several related works, more pertinent by Hitz *et al.**:

- User selection of Application ontology and instances
- Automatic transformation to a Target ontology for UI generation
- Proprietary annotations are required, limiting their universality



*M. Hitz, T. Kessel and D. Pfisterer, "Towards Sharable Application Ontologies for the Automatic Generation of UIs for Dialog based Linked Data Applications" 2017.

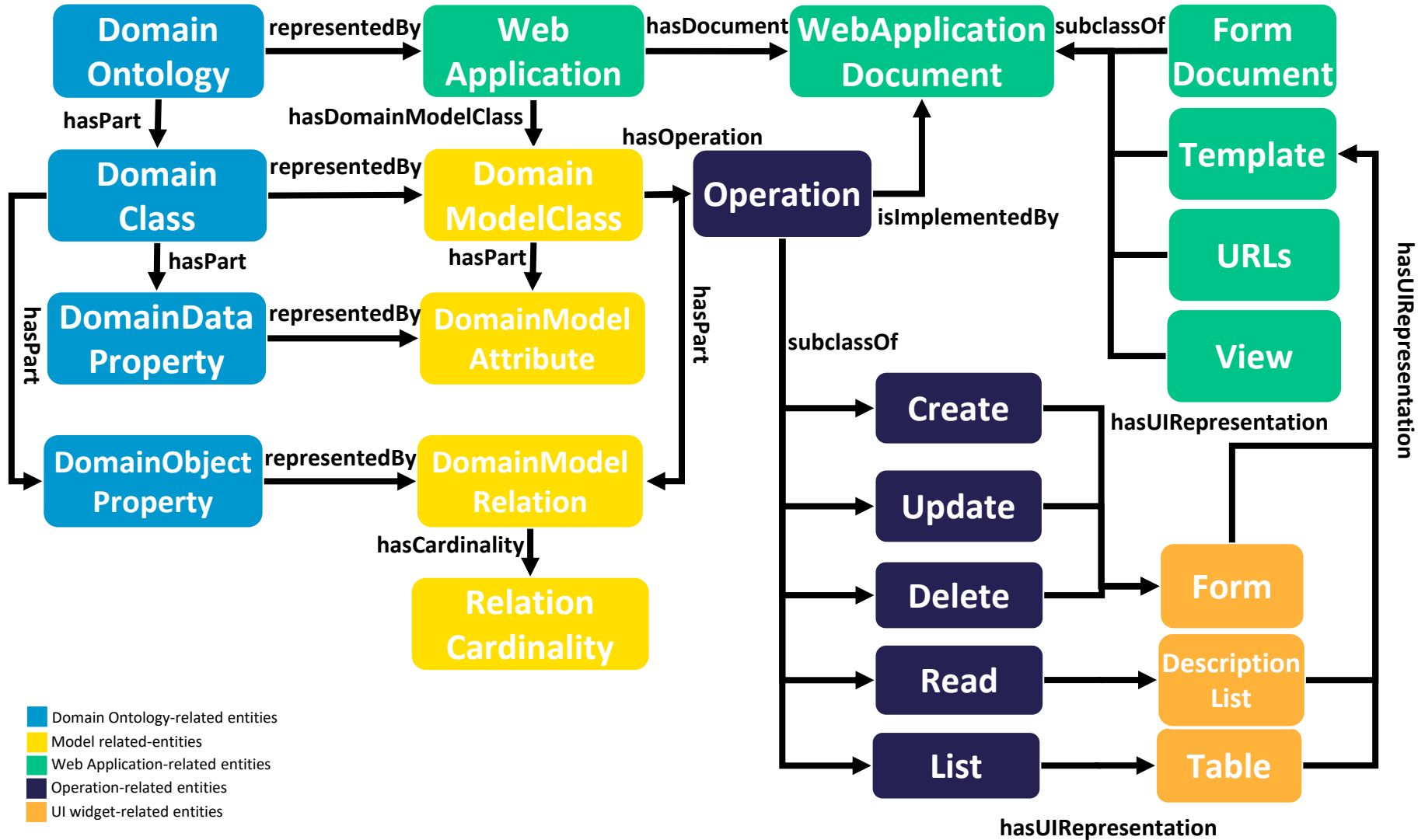
UI Ontologies

Existing ontologies **reused**:

- **Semantic UI** framework
- **LOV UI**, supported by the Linked Open Vocabularies (LOV)
- **RaUL**, RDFa User Interface Language (supported by LOV)
- **HUICA**, Hierarchical User Interface Component Architecture

Ontology	UI elements concepts (Forms, Tables, ...)	Style properties (color, background color, ...)	Data operations	Back-end concepts
Semantic UI	×	-	-	-
LOV UI	×	×	-	-
RaUL	×	-	×	-
HUICA	×	×	-	-

Ontology-based Web Application Ontology (OWAO)



Web Application Generator (GenAppi)

Loading domain and OWAO ontologies (OWL)

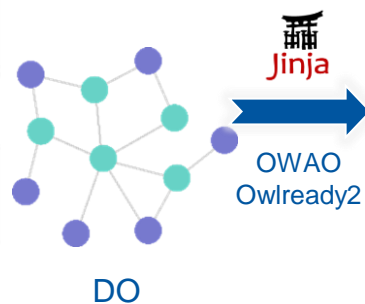
Transforming domain classes and object/data properties to Model classes and attributes

Generating Django view files for data operations (CRUD)

Creating Django Template and URLs files for the UI

Grouping files in specific directories

Migrating Model and server ready for start



```

class User(models.Model):
    hasRole=models.ManyToManyField("IrradiationFacilityRole")
    name=models.CharField(max_length = 50)
    surname=models.CharField(max_length = 50)
    email=models.CharField(max_length = 50)
    name=models.CharField(max_length = 50)
    def __str__(self):
        return self.name
    
```

Django Model



Django Templates



Web Application



DB or Triple Store

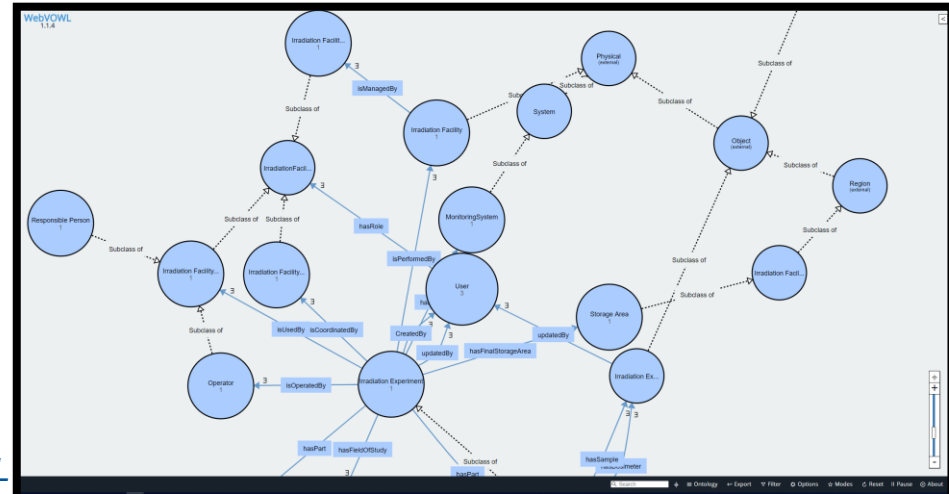


IEDM Use Case

- Input form for each domain-ontology class instance:
 - Data properties = input fields
 - Object properties = selection fields, links to the corresponding class forms
- Django authentication and authorisation**
- UI adaptable to user's preferences**
- Visualisation of domain ontology**

The screenshot shows the IEDM web application interface. On the left is a dark sidebar with a list of domain classes: ResponsiblePerson, SingularField, StorageArea, System, TechnicalRequirements, UnaryFunction, UnaryFunctionEntry, IrradiationExperiment, Requirements, DUTIrradiation, CumulatedQuantity, IrradiationExperimentObject, User, IrradiationFacilityRole, InteractionLength, InteractionLengthOccupancy, MonitoringSystem, DomainOfExperiment, StorageArea, IrradiationFacilityCoordinator, Operator, IrradiationFacility, and IrradiationFacilityManager. The main area is a form titled 'IEDM' with fields for 'UpdatedBy' (Blerina Giotse), 'CreatedBy' (Blerina Giotse), 'Height' (0.1), 'Identification' (SET-001029), 'Location' (Bld. 14 R12), 'Weight' (0.001), 'Width' (0.1), and 'Name' (SI-1029). There are 'Add' buttons next to the 'UpdatedBy' and 'CreatedBy' fields. At the bottom of the form are 'Cancel' and 'Create' buttons. On the right is a 'UI Preferences configurations' panel with options for 'Background body color', 'Font color', and 'Font size' (set to 18), and a 'Save' button.

Registering an IrradiationExperimentObject instance



Visualisation using WebVOWL

Generated Application vs. IDM

	IDM	Generated Application
Purpose	IRRAD facility data management	any domain
Software infrastructure	CERN	free and open-source
Storage	Oracle database	any relational database, ontology or triple store
Web Semantic technologies	no	yes
UI Customisation	no	yes
Functionalities	more advanced	CRUD* operations

CRUD* = Create Read Update Delete

User Interfaces Comparison: List Table

CERN Accelerating science Signed in as: blerina.gkotse@cern.ch Sign out Directory

HOME EXPERIMENTS BPM IRRAD INFO PREFERENCES

IRRAD Proton Facility AIDA 2020

IRRAD Data Manager

FCC-RADMON users

< Back Search... + New User

Name	Surname	E-mail	Telephone	Role	Actions
Blerina	Gkotse	blerina.gkotse@cern.ch	1111111	User	Edit Remove

IDM list table

- UnaryFunction
- UnaryFunctionEntry
- IrradiationExperiment
- Requirements
- DUTIrradiation
- CumulatedQuantity
- IrradiationExperimentObject
- User
- IrradiationFacilityRole
- InteractionLength

IEDM

User

[Create](#)

surname	email	name	hasRole	Actions
Gkotse	blerina.gkotse@cern.ch	Blerina	Responsible	View Update Delete

UI Preferences configurations

Background body color:

Font color:

Font size: px

[Save](#)

IEDM_APP list table

Outline

- Web Semantics
- Experimental Particle Physics
- Irradiation Experiment Data Management Ontology (IEDM)
- Automatic Generation of Web Applications from Ontologies
- **UI Personalization with new Ontology Embeddings**

User Experience

“User Experience (UX) refers to a person's entire experience using a particular product, system or service.”

©Interaction Design course UNIL



User Experience Honeycomb

©Peter Morville, “User Experience Design”
http://semanticstudios.com/user_experience_design/

User Experience

Bad UI design can cause serious human errors

The screenshot shows the Epic Hyperspace interface for a patient named Zztest, Ad. The patient's information is displayed in a yellow header bar, including MRN (18774711), DOB (4/15/1950), Age (60), Sex (M), Allergies (No Known Allergies), PCP (NO), Type (None), PH (BX35, HN35), Online (Basic), and Alerts (HM). A green arrow points to the Alerts field. The main content area is divided into several panels: Demographics (123 Easy St, 60 year old male), Problem List (ESOPHAGEAL REFLUX, ASTHMA NOS W/O STATUS ASTHM, ESSENTIAL HYPERTENSION NOS, ERRONEOUS ENCOUNTER), Health Maintenance (CREATININE, INFLUENZA VACCINE, LIPID SCREENING, PNEUMOCOCCAL VACCINE, POTASSIUM, TDAP VACCINE, UNIVERSAL HIV SCREENING DISCUSSION, VARICELLA ZOSTER VACCINE, COLORECTAL CANCER SCREENING), Allergies (No Known Allergies), Medications (PREVPAC Pack, lisinopril, tramadol, fluticasone, PREVPAC Pack, ranitidine), Immunizations (None), and Significant History/Details (Tobacco, Alcohol, 3 open orders, Language: UNKNOWN). The interface is cluttered with many tabs and buttons, and a large red sad face icon is overlaid on the bottom right.

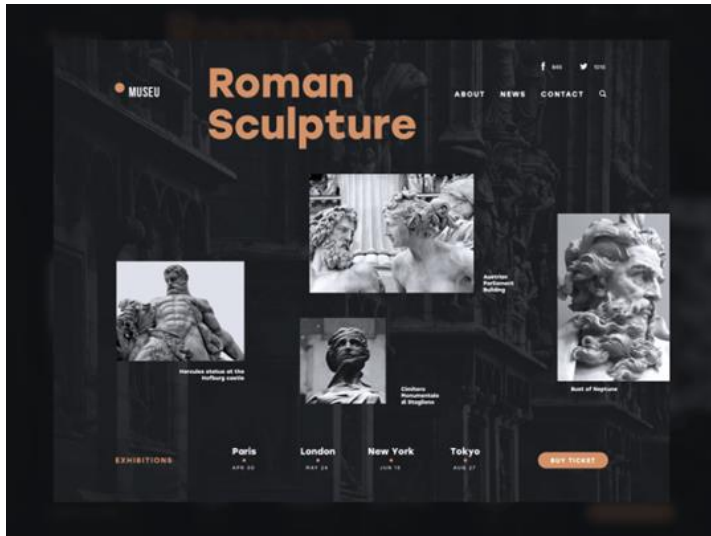
Epic software (in the USA)

<https://www.tragicdesign.com/>

Personalisation

User Experience factors:

- UI content
- Cultural background
- Physiology
- Cognitive capability



Images better displayed on black background

©Tubic Studio | Museu by Ernest Asanov



Eye tracking of websites for population of Western countries

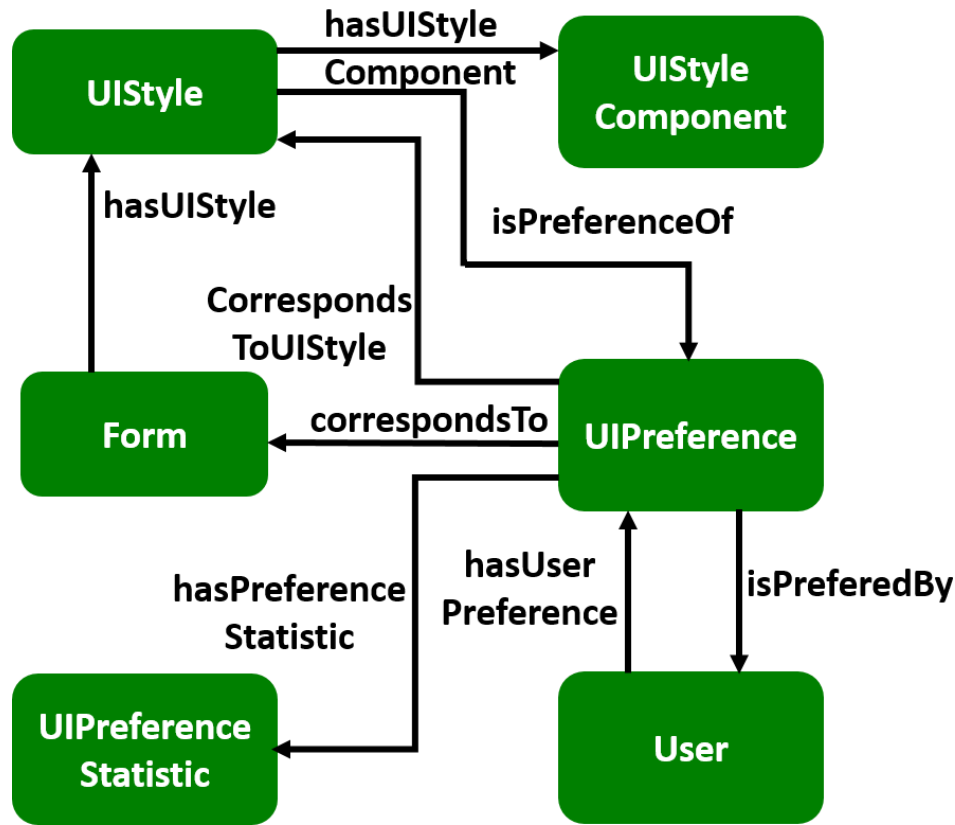
©nngroup.com NN/g



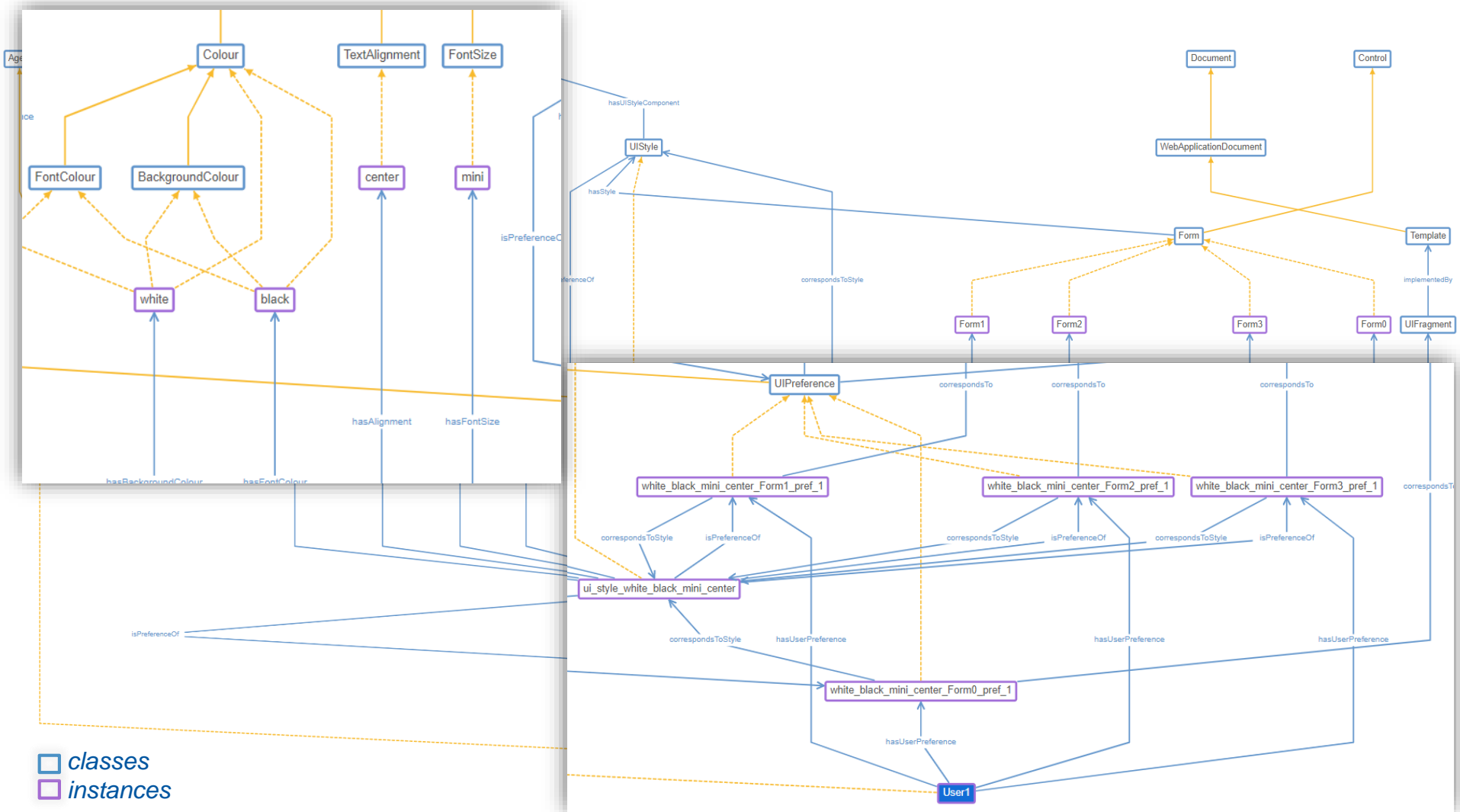
©iconfinder.com

OWAO UI Preferences

- Part of OWAO (Ontology-based Web Application Ontology)
- Used for recommending UI style



OWAO UI Preferences Example



Excerpt from OWAO ontology showing "User1" instance and UI Preferences (displayed from web Protégé)

Ontology Embedding

Ontology instances \rightarrow words, text

(e.g. *TemperatureSensor1* observes *Temperature*)

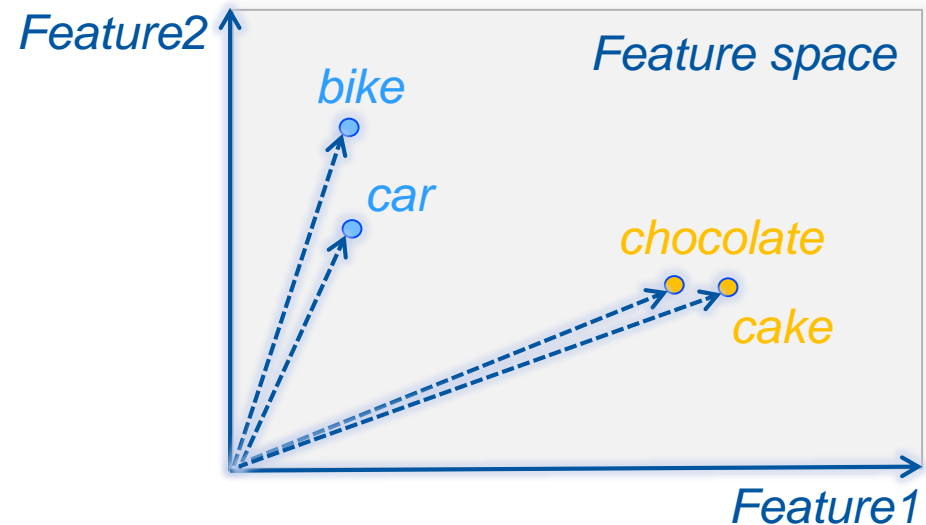
Word representation as feature vectors for tracking

- **similar words** and
- **words appearing in the same context**

Used in **recommender systems** for suggesting similar items

	<i>car</i>	<i>bike</i>	<i>cake</i>	<i>chocolate</i>
<i>Feature1</i>	1	1	6	7
<i>Feature2</i>	3	5	2	2
<i>Feature3</i>	4	4	1	1
<i>Feature4</i>	7	7	2	2

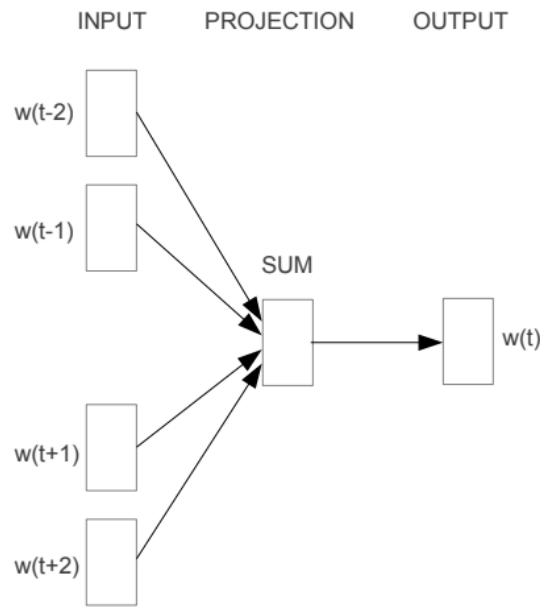
Feature Vectors



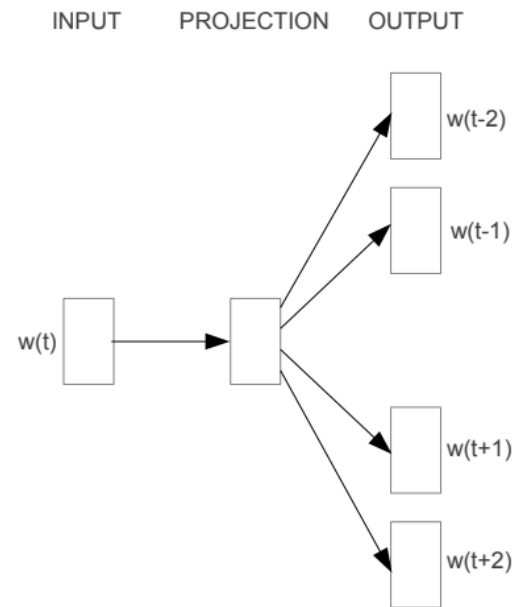
word2vec

NLP Model for generating word embeddings*:

- Continuous Bag of Words (CBOW)
- Skip-gram



CBOW



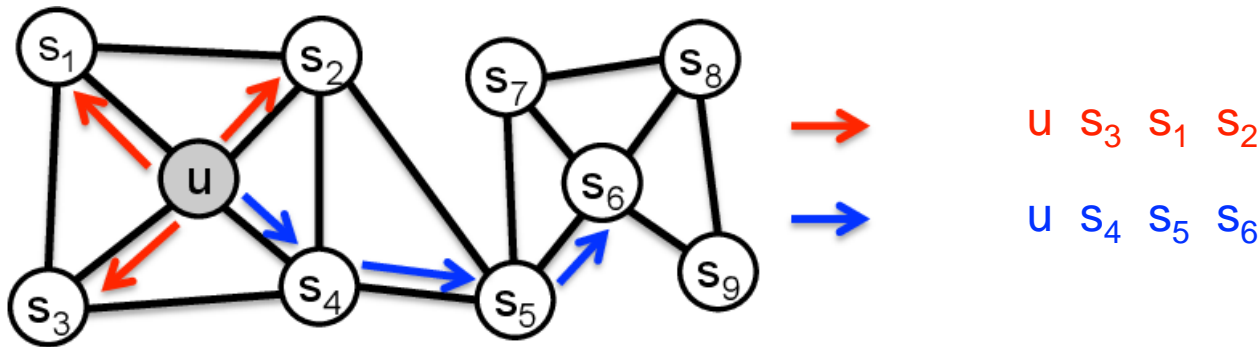
Skip-gram

$w = \text{word}$

*A. T. Mikolov et al., "Efficient Estimation of Word Representations in Vector Space"

node2vec / RDF2Vec

- Ontologies ~ directed graphs
- Random walks extraction to create “sentences”
- Two NLP models:
 - **node2vec** based on **Breadth-first Search (BFS)** and **Depth-first Search (DFS)**¹
 - **RDF2Vec** based on BFS of certain depth and Weisfeiler-Lehman (subtree comparison algorithm)²



BFS and **DFS** paths

$u \ s_3 \ s_1 \ s_2$

$u \ s_4 \ s_5 \ s_6$

¹A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks"

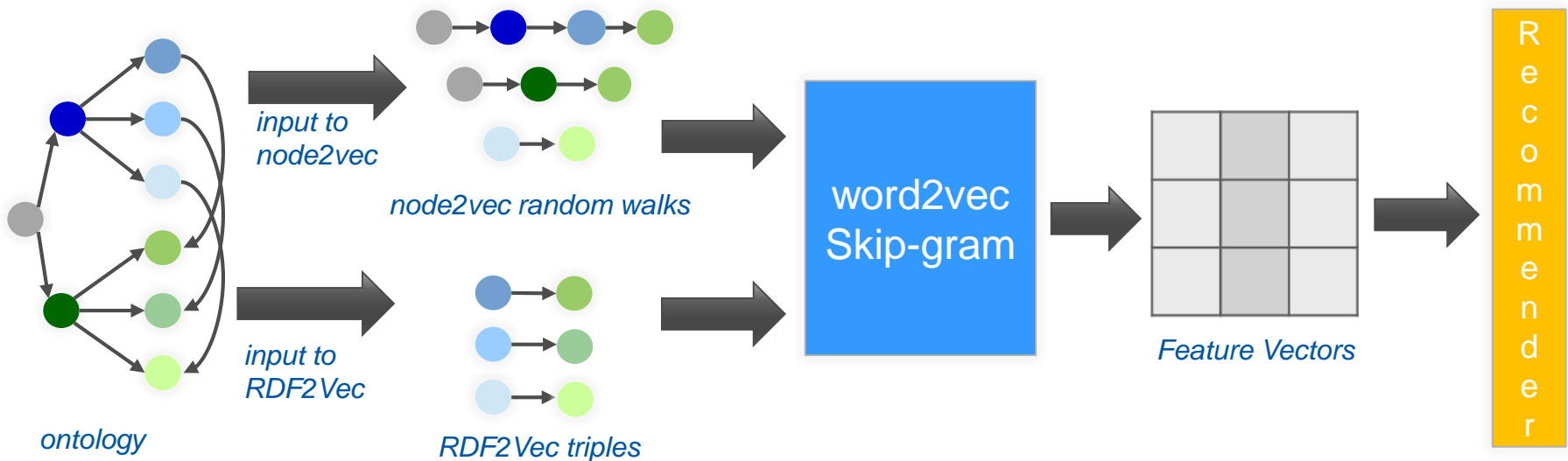
²P. Ristoski, et al., "RDF2Vec: RDF graph embeddings and their applications"

New embedding: *ontowalk2vec*

New hybrid NLP model:

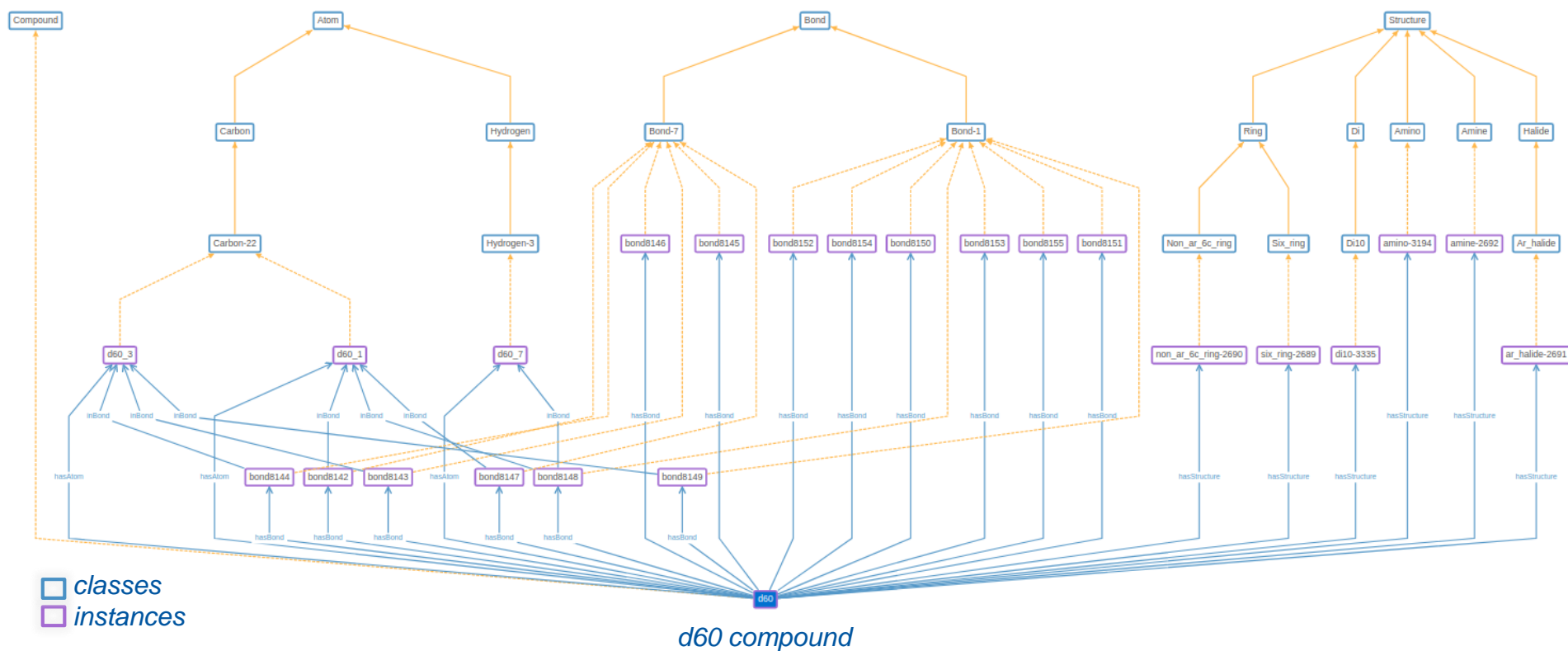
- *node2vec* → explicit structure of the ontology
- *RDF2Vec* → RDF triples

Algorithm:



Evaluation: MUTAG

- KB of 340 complex molecules (carcinogenic, i.e., “MUTAGenic” or not)
- Source ontology for classifying molecules (mutagenic / not-mutagenic)

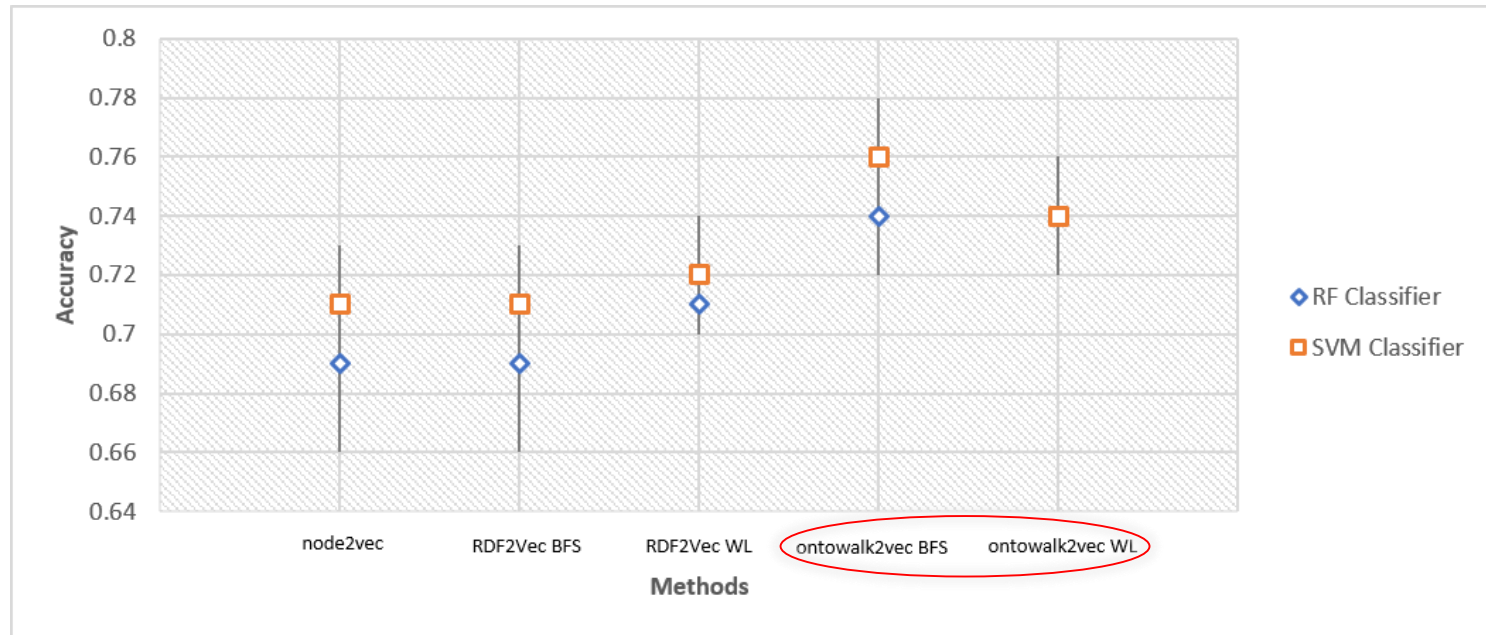


MUTAG Results

Experiments with *RDF2Vec*, *node2vec* and *ontowalk2vec* (run 10 times).

Classification accuracy test:

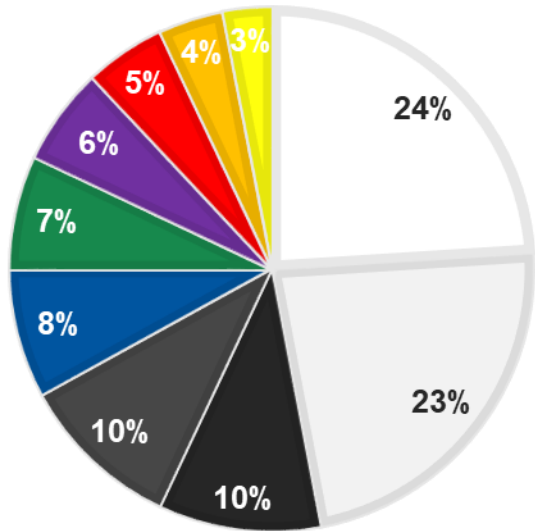
- **Random Forest (RF)**: decision trees on data sub-groups and averaging
- **Support Vector Machine (SVM)**: optimal hyperplane on labeled data



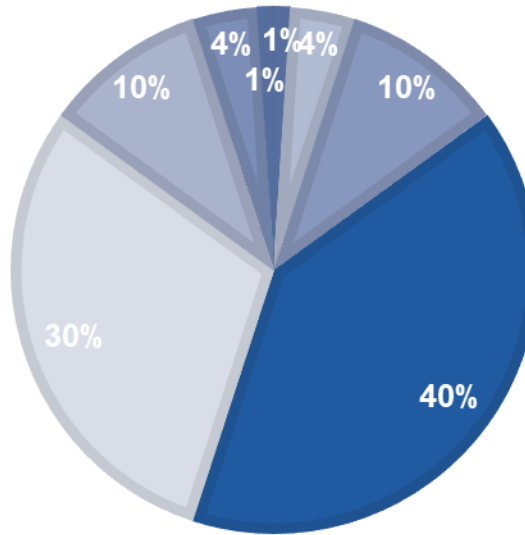
Evaluation: OWAO UI Preferences

- Lack of actual UI datasets → Artificial statistics-based data
- Four UI Style components taken as an example
- Statistics on UI preferences (bibliographic research)

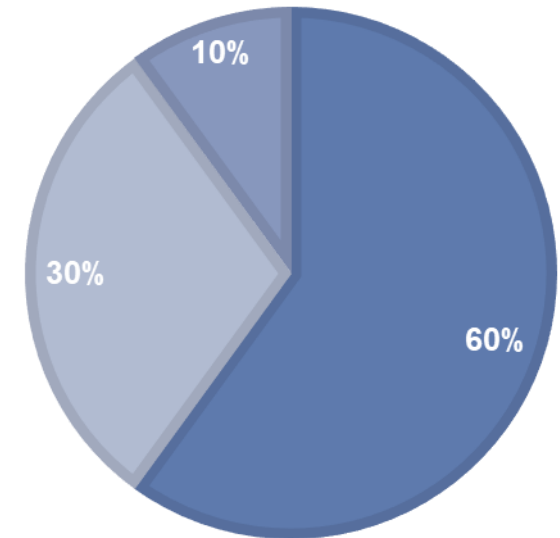
BACKGROUND-FONT COLOUR



FONT SIZE



TEXT ALIGNMENT



OWAO UI Preferences Results

Classification accuracy:

	Random Forest	SVM
BFS	1.00 (± 0.00)	1.00 (± 0.00)
WL	1.00 (± 0.00)	0.99 (± 0.01)

preferred / not preferred

	Random Forest	SVM
BFS	0.990 (± 0.001)	0.995 (± 0.002)
WL	0.993 (± 0.002)	0.995 (± 0.001)

popularity zero / low / medium / high

	Random Forest	SVM
BFS	0.83 (± 0.04)	0.94 (± 0.04)
WL	0.86 (± 0.06)	0.92 (± 0.05)

popularity low / medium / high

Cosine Similarity Testing

UpdatedBy
Blerina Gkotse

Add

CreatedBy
.....

Add

Height

Identification

Location

Weight

Width

Name

Cancel Create

Input instance

owao:ui_style_white_black_medium_left

UpdatedBy
Blerina Gkotse

Add

CreatedBy
.....

Add

Height

Identification

Location

Weight

Width

Name

Cancel Create

UpdatedBy
Blerina Gkotse

Add

CreatedBy
.....

Add

Height

Identification

Location

Weight

Width

Name

Cancel Create

font size

background

UpdatedBy
Blerina Gkotse

Add

CreatedBy
.....

Add

Height

Identification

Location

Weight

Width

Name

Cancel Create

text alignment

Output instances

Instances	Cosine Similarity
<i>owao:ui_style_white_black_medium_center</i>	0.956
<i>owao:ui_style_light_grey_black_medium_left</i>	0.955
<i>owao:ui_style_light_grey_black_large_left</i>	0.941
<i>owao:ui_style_white_black_large_left</i>	0.931
<i>owao:ui_style_light_grey_black_medium_center</i>	0.924

Similar instances using ontowalk2vec BFS

Conclusion

- **Interdisciplinary work** bridging the gap between Web Semantics and EPP
- **Automatic generation** of web applications **from ontologies** and KGs:
 - GenAppi methodology and software tools developed
 - Use case based on the IEDM ontology (and other ontologies)
- Generated web applications:
 - Based on **web semantic technologies**, enabling data integration
 - **Functional** (as is) or **starting point** (for further software development)
- Ontology embedding (ontowalk2vec):
 - Evaluated with two ontologies
 - **Some evidence of better accuracy** than state of the art
 - Generating OWAO **embeddings** and **finding similarities in instances**
 - Used for **recommending personalized UIs**

Perspectives

- IEDM:
 - Further **validation** by the EPP community
 - **Standardisation** of data management for irradiation experiments
 - **Common tools** in the EPP community
- GenAppi methodology:
 - **More UI components** to be integrated in OWAO
- ontowalk2vec:
 - **Collection of data** regarding actual UI preferences
 - Evaluation with **larger datasets**
 - **Extended use** (irradiation experiments similarities – IEDM)

Ontology-based Generation of Personalised Data Management Systems: an Application to Experimental Particle Physics

PhD thesis defense of Blerina GKOTSE

Blerina.Gkotse@cern.ch

MINES ParisTech, PSL University, France

25 September 2020

Jury:

Laura GONELLA, Rapporteur

Jean-Baptiste LAMY, Rapporteur

Laurent DUSSEAU, Examineur

Theodora VARVARIGOU, Examineur

Pierre JOUVELOT, Directeur de thèse

Federico RAVOTTI, Maître de thèse

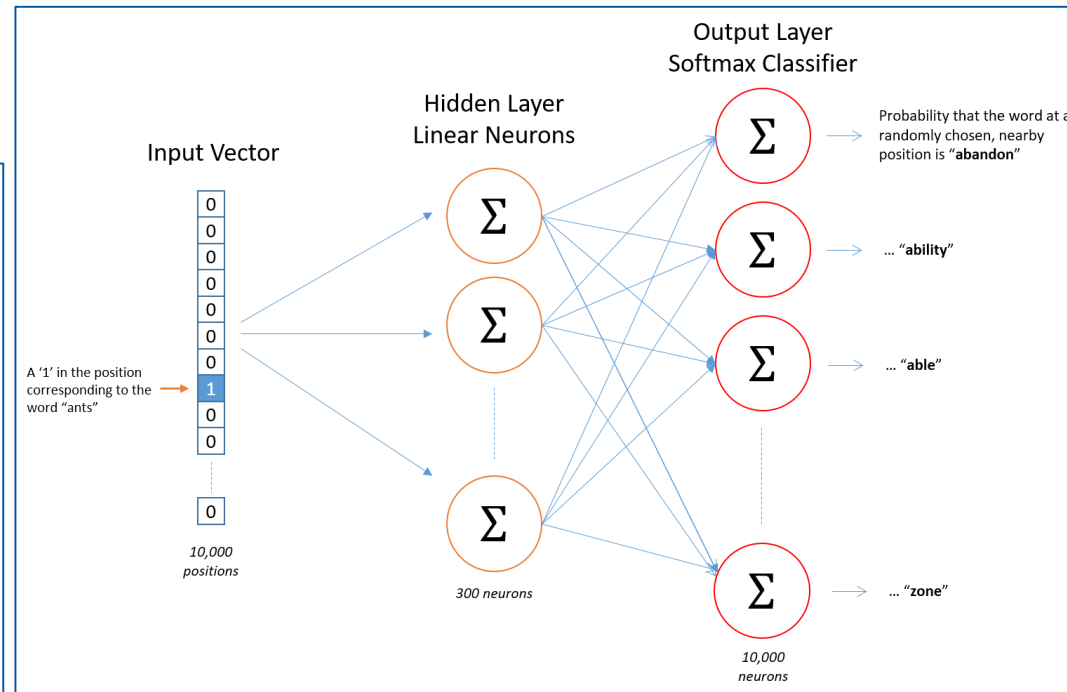
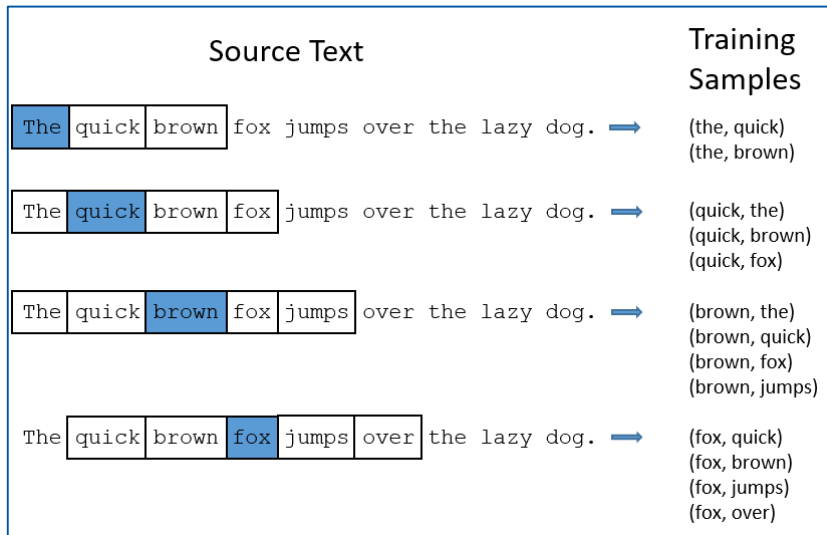
This work has received funding from the European Union's Horizon 2020 Research and Innovation program under Grant Agreement no. 654168.



Backup: word2vec - Skip-gram

Given a sequence of training words $w_1, w_2, w_3, \dots, w_T$, maximize log probability:

$$\frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t)$$



Example for training word sentences*

Word2Vec Neural Network*

*McCormick, C. (2016, April 19). *Word2Vec Tutorial - The Skip-Gram Model*. Retrieved from <http://www.mccormickml.com>

Backup: NLP Model node2vec

Ontologies considered as directed graphs → need for extracting **random walks** in the graph to create “sentences”.

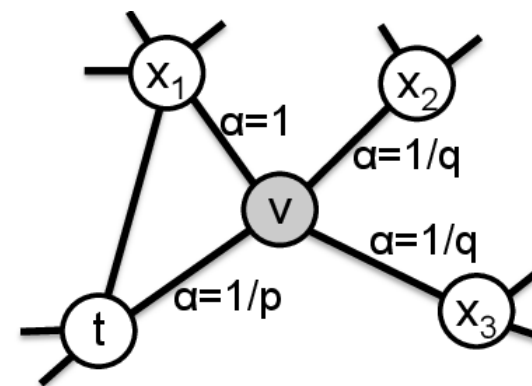
- Two algorithms for traversing graphs:
 - Breadth-first Search (BFS)
 - Depth-first Search (DFS)
- Based on hyperparameters:
 - p : controlling the probability that a node in the walk is revisited
 - q : controlling the choice whether a BFS or DFS approach should be employed
- word2vec for generating the final embeddings

$$P(c_i = x \mid c_{i-1} = v) = \begin{cases} \frac{\pi_{v,x}}{Z} & \text{if } (v, x) \in E \\ 0 & \text{otherwise} \end{cases}$$

$\pi_{v,x}$ is the unnormalized transition probability between nodes v and x , and Z is the normalizing constant.

$$\alpha_{pq}(t, x) = \begin{cases} \frac{1}{p} & \text{if } d_{tx} = 0 \\ 1 & \text{if } d_{tx} = 1 \\ \frac{1}{q} & \text{if } d_{tx} = 2 \end{cases}$$

$\pi_{v,x} = \alpha_{pq}(t, x) \cdot w_{v,x}$
 d_{tx} = shortest path distance between t and x



The walk transitions from t to v and is evaluating its next possible steps out of node v

*A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks"

Backup: Hyperparameters

Hyperparameter	Explanation	Default	Range	Comments
alpha	The initial learning rate	0.025	[0.1, 0.05, 0.025, 0.01]	Too big learning rate may not converge, too small may get to a local minimum
iteration	Number of iterations (epochs) over the corpus	5	[1, 5, 20, 50]	The more iteration, the better the model is trained
vector size	Dimensionality of the word vectors	100	[20, 100, 200, 500]	Bigger vector space the more features are learned for each word
window	Maximum distance between the current and predicted word within a sentence	5	[3, 5, 7]	since we are using RDF triples it would be wise to keep a window of min 3 words, and we add 2 more levels about the preferences and styles
min_count	Ignores all words with total frequency lower than this.	5	[0, 1]	Since there are not many data and repeating instances, we don't want to remove too many data
negative / hierarchical softmax	If negative > 0, negative sampling will be used, the int for negative specifies how many "noise words" should be drawn (usually between 5-20). If set to 0, no negative sampling is used. If negative = 0, hierarchical softmax is used	5	[0, 1, 5, 10]	Since we don't have too many noisy words, we try to run our algorithm with low negative sampling
depth	maximum level traversing the graph	1	[1, 2, 3]	After 3, too many and long random walks generated and the training becomes too heavy

Backup: Hyperparameters

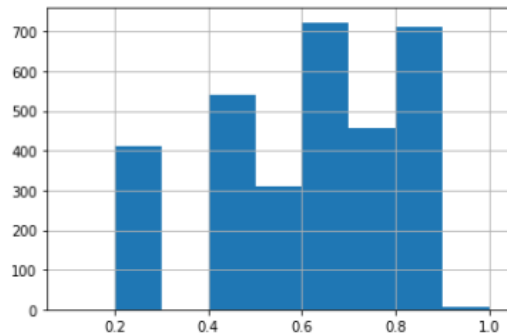


Figure 6.12: BFS F1 score histogram

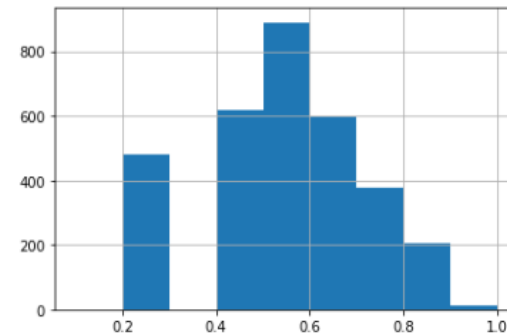


Figure 6.13: WL F1 score histogram

win	vs	Hyperparameters						Evaluation metrics		
		hs	lr	iter	neg	cnt	dep	rec	prec	F1
3	200	1	0.010	1	0	1	1	1	0.85	0.923
7	100	0	0.025	1	5	0	2	1	0.857	0.923
5	500	0	0.025	1	20	0	2	1	0.857	0.923
7	500	0	0.025	1	10	0	2	1	0.857	0.923
5	200	1	0.010	1	0	1	1	1	0.857	0.923
5	100	10	0.025	1	5	0	2	1	0.857	0.923
7	500	0	0.010	5	5	1	1	1	0.857	0.923
7	20	0	0.025	50	1	0	2	0.833	1	0.909

Table 6.19: Top F1 score (> 0.9) for BFS

win	vs	Hyperparameters						Evaluation metrics		
		hs	lr	iter	neg	cnt	dep	rec	prec	F1
7	100	0	0.010	5	5	1	1	1	0.857	0.923
7	500	0	0.025	1	10	0	2	1	0.857	0.923
3	500	0	0.050	1	5	0	1	1	0.857	0.923
5	500	0	0.025	1	20	1	1	1	0.857	0.923
5	20	0	0.025	1	20	1	1	1	0.857	0.923
7	20	0	0.025	1	10	1	1	1	0.857	0.923
7	100	0	0.025	1	10	0	2	1	0.857	0.923
3	100	0	0.050	1	5	0	1	1	0.857	0.923
7	20	0	0.010	5	5	1	1	1	0.857	0.923
3	20	0	0.050	1	5	1	1	1	0.857	0.923

Table 6.20: Top F1 score (> 0.9) for WL

Backup: Metrics

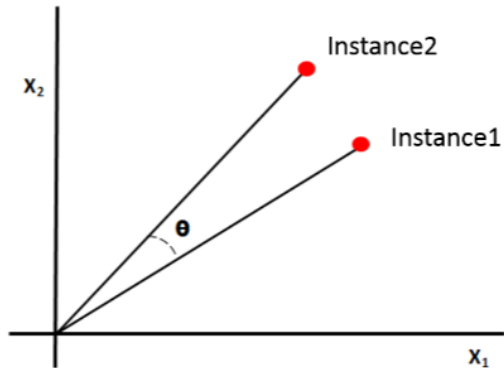
		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

$$\text{Precision} = \frac{||\text{True Positives}||}{||\text{True Positives}|| + ||\text{False Negatives}||}$$

$$\text{Recall} = \frac{||\text{True Positives}||}{||\text{True Positives}|| + ||\text{False Positives}||}$$

$$\text{F1 score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Backup: Cosine Similarities Example

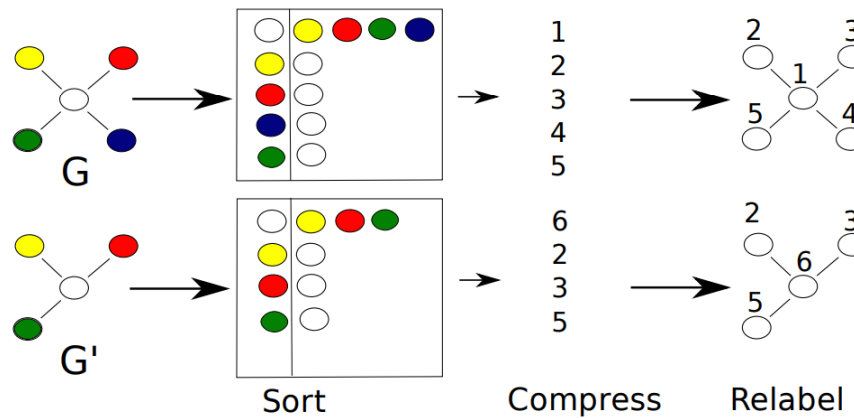
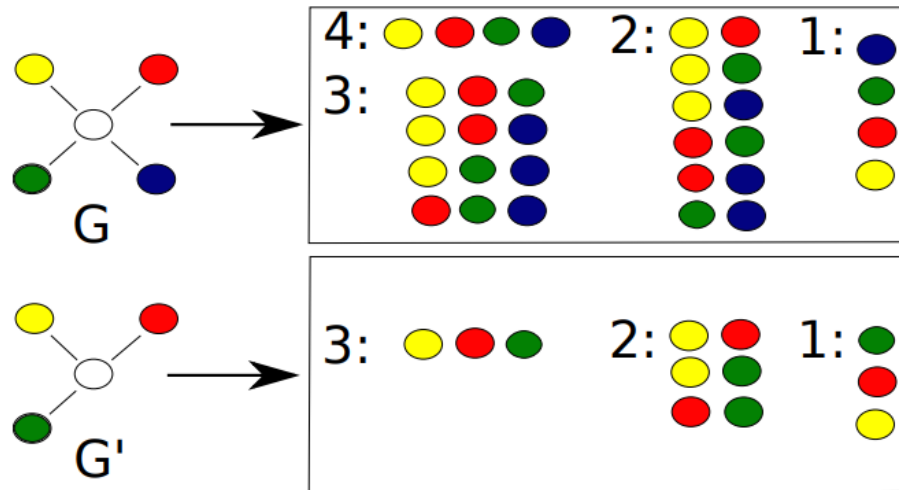


$$\cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$

$$\begin{bmatrix} [1. & 0.75 & 0.75 & \dots & 0.25 & 0. & 0.] \\ [0.75 & 1. & 0.75 & \dots & 0. & 0.25 & 0.] \\ [0.75 & 0.75 & 1. & \dots & 0. & 0. & 0.25] \\ \dots \\ [0.25 & 0. & 0. & \dots & 1. & 0.75 & 0.75] \\ [0. & 0.25 & 0. & \dots & 0.75 & 1. & 0.75] \\ [0. & 0. & 0.25 & \dots & 0.75 & 0.75 & 1.] \end{bmatrix}$$

Cosine similarity table

Backup: Weisfeiler Lehman Algorithm



Backup: Recommenders

- Content-based methods:

A model for user-item interactions where items or users' explicit features are provided

+ **Advantages:** User independence, transparency, no cold start

- **Disadvantages:** Overspecialisation

- Collaborative filtering methods:

- Model-based: A model for user-item interactions where users and items representations are learned from interaction matrix

- Memory-based: Depending on similarities between users or items in terms of observed interactions

+ **Advantages:** It may suggest more different items

- **Disadvantages:** It requires a lot of interactions, users and item ratings

Backup: GenAppi Algorithm

Algorithm 1 GenAppi

Input: *domain_ontology*

Output: client code and server

```
1: load domain_ontology, owao
2: load Jinja_templates
3: map_domain_ontology_to_model(domain_ontology, owao)
4: create URL_file
5: for op in owao.Operations() do
6:   create view_file(op)
7:   get template(op) from Jinja_templates
8:   for cl in domain_ontology.classes() do
9:     adjust template(op) to cl
10:    view_file(op).append(cl.template(op))
11:    create cl.URL(op)
12:    URL_file.append(cl.URL(op))
13:  end for
14: end for
15: WebVOWL_JSON_file ← owl2vowl(domain_ontology)
16: assemble files in directory
17: migrate Model
18: start server
```

Backup: *ontowalk2vec* Algorithm

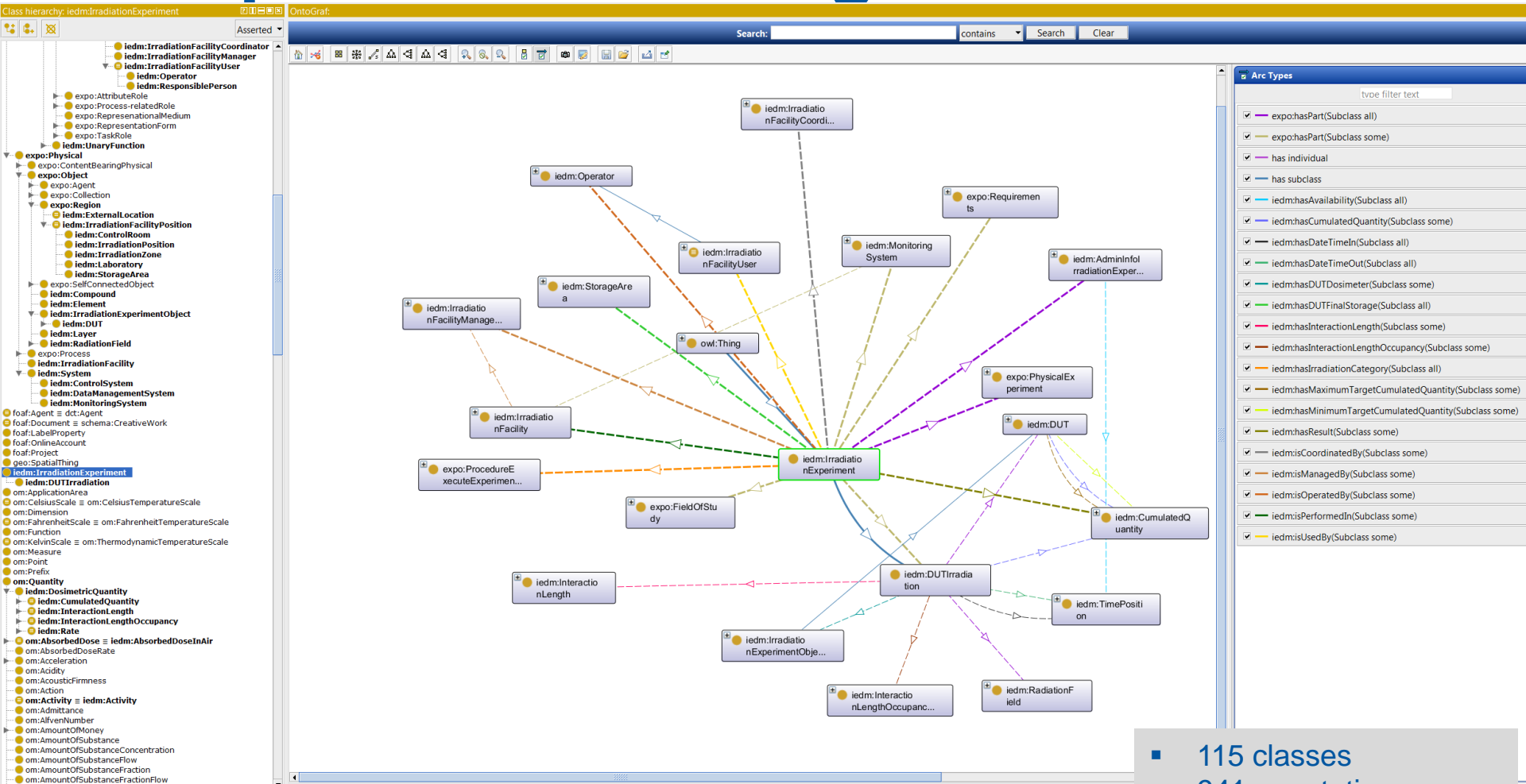
Algorithm 2 *ontowalk2vec* and classification

Input: *ontology*, *training_data*, *test_data*

Output: ontology embeddings and classification analysis data

```
1: Read ontology, training_data, test_data
2: Separate training_data to training_members and training_labels
3: Separate test_data to test_members and test_labels
4: graph  $\leftarrow$  ontology_to_graph_conversion(ontology)
5: node2vec_graph  $\leftarrow$  node2vec.graph(graph, directed, p, q)
6: node2vec_graph.preprocess_transition_probabilities()
7: node2vec_walks  $\leftarrow$  node2vec_graph.simulate_walks()
8: bfs_walks  $\leftarrow$  RDF2VecBFS(graph, training_members  $\cup$  test_members)
9: bfs_embeddings  $\leftarrow$  word2vec(bfs_walks  $\cup$  node2vec_walks)
10: wl_walks  $\leftarrow$  RDF2VecWL(graph, training_members  $\cup$  test_members)
11: wl_embeddings  $\leftarrow$  word2vec(wl_walks  $\cup$  node2vec_walks)
12: for embeddings in [bfs_embeddings, wl_embeddings] do
13:   Separate embeddings to training_embeddings and test_embeddings
14:   rf  $\leftarrow$  RandomForestClassifier()
15:   svm  $\leftarrow$  SVM()
16:   for classifier in [rf, svm] do
17:     classifier.fit(training_embeddings, training_labels)
18:     classifier.predictions  $\leftarrow$  classifier.predict(test_embeddings)
19:     classifier.accuracy  $\leftarrow$ 
20:       classifier.accuracy_score(test_labels, predictions)
21:     classifier.confusion_matrix  $\leftarrow$ 
22:       confusion_matrix(test_labels, classifier.predictions)
23:   end for
24: tsne  $\leftarrow$  TSNE(embeddings).fit_transform(embeddings)
25: Plot tsne
26: top_similarities  $\leftarrow$  embeddings.most_similar(instance1)
27: similarity_value  $\leftarrow$  embeddings.similarity(instance1, instance2)
28: end for
```

Backup: IEDM in Protégé



- 115 classes
- 941 annotations
- 24 object properties
- 16 data properties