PERICLES - Promoting and Enhancing Reuse of Information throughout the Content Lifecycle taking account of Evolving Semantics

[Digital Preservation]

DELIVERABLE 4.5

CONTEXT-AWARE CONTENT INTERPRETATION

GRANT AGREEMENT: 601138

SCHEME FP7 ICT 2011.4.3

Start date of project: 1 February 2013

Duration: 48 months

Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)		
Dissemination level		
PU	PUBLIC	X
PP	Restricted to other PROGRAMME PARTICIPANTS (including the Commission Services)	
RE	RESTRICTED to a group specified by the consortium (including the Commission Services)	
CO	CONFIDENTIAL only for members of the consortium (including the Commission Services)	

Revision History

V #	Date	Description / Reason of change	Author
V0.1	01.06.16	Outline and first draft	HB, CERTH
V0.2	20.07.16	Outline and second draft	HB, CERTH
V0.3	01.08.16	Pre-final version of Chapter 5	CERTH
V0.4	10.08.16	Pre-final version of Chapters 3-4	HB
V0.5	31.08.16	Final draft	HB, CERTH
V1.0	05.09.16	Final version submitted to internal reviewers	HB, CERTH
V1.1	25.09.16	Final version submitted to KCL	HB, CERTH

Authors and Contributors

Authors

Partner	Name
HB	S. Darányi, P. Wittek
CERTH	E. Kontopoulos, K. Konstantinidis, M. Riga, P. Mitzias, T. Stavropoulos, S. Andreadis, A. Maronidis, A. Karakostas, S. Tachos, V. Kaltsa, M. Tsagiopoulou, K. Avgerinakis.

Reviewers

Partner	Name
XRCE	J.-Y. Vion-Dury
UGOE	D. Kurzawe

Table of Contents

GLOSSARY 6

1. EXECUTIVE SUMMARY 9

2. INTRODUCTION & RATIONALE 10

2.1. What to expect from this Document 10
2.2. Relation to other Work Packages 11
2.3. Relation to other WP4 Tasks 11
2.4. Document Structure 12

3. CONTEXTUALISED CONTENT INTERPRETATION 13

3.1. Overview & Motivation 13
3.2. Evolving Semantics & Physics as its Metaphor 14
3.3. Semantic Reasoning for DP 18
3.4. Chapter Summary 18

4. QUANTUM-LIKE ANALYSIS FOR CONTEXTUAL CONTENT INTERPRETATION 19

4.1. Quantum-like Methods for Text Interpretation 20
4.2. Investigation into the Quantum-like Nature of Semantic Content Related User Behaviour 22
4.2.1. Eye Gaze Fixation during Information Searching & Contextuality 22
4.2.2. Citation Behavior and Entanglement 28
4.3. Description and Implementation of Dual Content Representations and Development of Quantum-based Models for Semantic Content Classification 35
4.3.1. Particle-like Index Terms, Drifts and "Gravity" 35
4.3.2. Index Terms Between Classical and Quantum Mechanics 45
4.3.3. Index Term Drifts and Entanglement: Integrating Two Analytical Approaches 51
4.3.4. Outlines of an Energy Regime for Word and Sentence Semantics 53
4.4. Chapter Summary 54

5. SEMANTIC REASONING FOR CONTEXTUAL CONTENT INTERPRETATION 56

5.1. Background 56
5.1.1. Description Logics 56
5.1.2. Semantic Reasoning and DL Reasoning Services 57
5.2. PERICLES Semantic Interpretation Framework 58
5.2.1. Representing Content, Context & Use-context 59
5.2.2. Ontological Inference 60
5.2.3. SPIN Reasoning Layer 66
5.2.4. Using Background Domain Knowledge 73
5.2.5. Contextualised Reasoning on Semantic Drifts 74
5.3. Uncertainty Handling 80
5.3.1. Approaches for Handling Inconsistent & Missing Knowledge 80
5.3.2. Rule-based Uncertainty Management 82
5.3.3. An Uncertainty Management Example: Impacted & Unimpacted DOs 82
5.3.4. Other Uncertainty Management Applications in PERICLES 86
5.4. Chapter Summary 86

6. CONCLUSIONS AND NEXT STEPS 87
6.1. CONCLUSIONS 87
6.2. NEXT STEPS 87
7. REFERENCES 89

Glossary

Abbreviation / Acronym	Meaning
A&M	Arts and Media
AA	Advanced Access
AHCI	Arts and Humanities Citation Index
ANN	Artificial Neural Network
AOI	Area of Interest
API	Application Program Interface
BDA	Born-Digital Archives
BMU	Best Matching Unit
BOSS	Baryon Oscillation Spectroscopic Survey
DL	Description Logics
DNA	Deoxyribonucleic Acid
DO	Digital Object
DoW	Description of Work
DVA	Digital-Video Artworks
CDS	Compositional Distributional Semantics
CM	Classical Mechanics
DEM	Digital Ecosystem Model
DL	Description Logics
DP	Digital Preservation
DQC	Dynamic Quantum Clustering
EEG	Electroencephalogram
EM	Electromagnetism
EP	External Potential
ES	Evolving Semantics
ESOM	Emergent Self-Organizing Maps
GloVe	Global Vector for Word Representation
GTR	General Theory of Relativity
GVSM	Generalized Vector Spaces Model

HCA	Hierarchical Cluster Analysis
IP	Interaction Potential
IR	Information Retrieval
ISAD(G)	General International Standard Archival Description
JCR	Journal Citation Reports
JSON	JavaScript Object Notation
KE	Kinetic Energy
LIS	Library and Information Studies
LOD	Linked Open Data
LRM	Linked Resource Model
LSA	Latent Semantic Analysis
LSTE	Long Short-Term Memory
LTDP	Long Term Digital Preservation
MEDLINE	Medical Literature Analysis and Retrieval System Online
MeSH	Medical Subject Headings of the National Library of Medicine
ML	Machine Learning
Ncpol2spda	A software tool for non-commuting polynomial optimization problems by semidefinite programming
OAIS	Open Archival Information System
ODP	Ontology Design Pattern
OHSUMED	Oregon Health Sciences University Medical Test Collection
ORSD	Ontology Requirements Specification Document
OS	Operating System
OWL	Web Ontology Language
PCA	Principal Component Analysis
PSI	Predication-based Semantic Indexing
QA	Quality Assurance
QC	Quantum Clustering
QFT	Quantum Field Theory
QL	Quantum-like(ness)
QM	Quantum Mechanics

QML	Quantum Machine Learning
QT	Quantum Theory
RBF	Radial Basis Function
RDF	Resource Description Framework
SBA	Software-Based Artworks
SERP	Search Engine Results Page
SDP	Semidefinite Programming
SDSS	Sloan Digital Sky Survey
Somoclu	Self-Organizing Maps Over a Cluster
SOM	Self-Organizing Map
SOS	Sum-of-squares (decomposition)
SPARQL	SPARQL Protocol and RDF Query Language
SPIN	SPARQL Inferencing Notation
SSCI	Social Sciences Citation Index
SVD	Singular Value Decomposition
SVM	Support Vector Machines
SWRL	Semantic Web Rule Language
TC	Text Categorization
tfidf	Term frequency x inverse document frequency (formula)
VSA	Vector Symbolic Architectures
VSM	Vector Space Information Retrieval Model
WoS	Web of Science

1. Executive Summary

The current deliverable summarises the work conducted within task T4.5 of WP4, presenting our proposed approaches for contextualised content interpretation, aimed at gaining insightful contextualised views on content semantics. This is achieved through the adoption of appropriate context-aware semantic models developed within the project, and via enriching the semantic descriptions with background knowledge, deriving thus higher level contextualised content interpretations that are closer to human perception and appraisal needs.

More specifically, the main contributions of the deliverable are the following:

- A theoretical framework using physics as a metaphor to develop different models of evolving semantic content.
- A set of proof-of-concept models for semantic drifts due to field dynamics, introducing two methods to identify quantum-like (QL) patterns in evolving information searching behaviour, and a QL model akin to particle-wave duality for semantic content classification.
- Integration of two specific tools, Somoclu for drift detection and Ncpol2spda for entanglement detection.
- An “energetic” hypothesis accounting for contextualized evolving semantic structures over time.
- A proposed semantic interpretation framework, integrating (a) an ontological inference scheme based on Description Logics (DL), (b) a rule-based reasoning layer built on SPARQL Inference Notation (SPIN), (c) an uncertainty management framework based on non-monotonic logics.
- A novel scheme for contextualized reasoning on semantic drift, based on LRM dependencies and OWL’s punning mechanism.
- An implementation of SPIN rules for policy and ecosystem change management, with the adoption of LRM preconditions and impacts. Specific use case scenarios demonstrate the context under development and the efficiency of the approach.
- Respective open-source implementations and experimental results that validate all the above.

All these contributions are tightly interlinked with the other PERICLES work packages: WP2 supplies the use cases and sample datasets for validating our proposed approaches, WP3 provides the models (LRM and Digital Ecosystem models) that form the basis for our semantic representations of content and context, WP5 provides the practical application of the technologies developed to preservation processes, while the tools and algorithms presented in this deliverable can be deployed in combination with test scenarios, which will be part of the WP6 test beds.

2. Introduction & Rationale

According to [Schlieder, 2010], Long Term Digital Preservation (LTDP) should be concerned with three kinds of changes, all influencing the future accessibility of preserved content: **technology drifts**, **semantic drifts** and **social value shifts** – including attitudes, preferences etc. – with an impact on interpretation and reasoning. Because the question is not *if* but *when* a next technology will affect LTDP [Wang & Gai, 2014], under pressure from the inevitable we address the latter two drift types in this deliverable.

Examples for already happening drastic technology changes include e.g. the use of DNA for very long term digital preservation (according to Grass et al., in the range of 2000 years [Grass et al., 2015]; combined with nanostructured glass storage, for 13.8 billion years [Kazansky et al., 2016]). Parallel efforts have paved the way for this breakthrough lately [Church et al., 2012; Goldman et al., 2013; Bornholt et al., 2016]. The second example is contemporary Google, who frame themselves as an artificial intelligence (AI) company instead of an information retrieval (IR) one, and aim to combine deep learning with symbolic approaches to reasoning.

Because Digital Preservation (DP) does not exist in a vacuum, such pace of progress necessitates that:

1. We look into advanced methods to describe and represent evolving semantics (ES); and
2. Indicate novel opportunities how its nature invites new approaches to argumentation.

In this context, and summarising the work of task T4.5, this deliverable presents our proposed approaches for **addressing context-aware content interpretation**.

2.1. What to expect from this Document

In this deliverable, we shall investigate the view that, whereas LTDP is facing the problem of an unknown deadline[1], convergence between increasingly scalable content and growing DP needs necessitates new and adequate computational solutions. As LTDP has to handle increasing amounts of data/knowledge, enabling computationally efficient and advanced access to it will be one of the key issues in the future.

- As a corollary, we assume that wherever multivariate statistics is involved in creating semantic spaces, *distributional semantics*, that is, the theory that the meaning of the words derives from their distributional patterns, will remain a crucial ingredient for experimentation, products and services. On the other hand, ontology-based solutions will continue to rely on *logical semantics*. We expect research to come up with converging approaches between these two tracks.
- To model the dynamics of evolving content, in Chapter 4 we will use the metaphor of "content as energy" to help statistical model building. This goes back to two reasons. The first is that the concept of dynamics is coupled with the notion of energy in physics, so it makes sense to test imported methodology as an interdisciplinary research direction to study the behaviour of content over time. Secondly, information and knowledge are stored in structures, being constantly reconfigured as a result of intellectual progress, where this progress is a function of work investment. Conveniently, the energy metaphor includes the concept of work too, leaving room for new, improved models of evolving semantic content.

[1] As the IBM Research Labs in Israel states: "*Today's society is facing the Digital Dark Age: as the world becomes digital, the world's data is in increasing danger of being lost. (…) LTDP is particularly challenging when preserving large amounts of heterogeneous data for very long periods of time of tens or even hundreds of years.*" [Source: http://research.ibm.com/haifa/projects/storage/ltdp/index.shtml]

In the overlap where information representation meets scalable new computational methods, active research is going on to identify the kind of semantics fitting the formalism of quantum mechanics [Bruza & Busemeyer, 2012; Heunen et al., 2013], and also actual physical systems in quantum computing or quantum information theory [Zeng & Coecke, 2016]. The so far most successful contender, **compositional distributional semantics**, combines logical and distributional semantics, and addresses both word and sentence meaning. Also, context-dependence is a hallmark feature of quantum physics, and quantum theory offers an extensive mathematics toolbox to deal with the phenomenon: we believe that it was inevitable to bridge the disciplines and address contextuality through methods that were otherwise entirely absent from DP. On the other hand, the representation of content and context semantics via ontological structures in PERICLES, allows us in Chapter 5 to deploy **powerful semantic inference and reasoning techniques** [McGuinness & Da Silva, 2004]. These include both first order monotonic and nonmonotonic logics, which can be applied in a range of scenarios, like e.g. (a) deriving implicit knowledge from explicitly asserted information, (b) allowing different interpretations of content based on contextual information, and (c) assisting preservation experts in determining DP-related risks and respective mitigation actions.

Our effort is significant for LTDP because of long-term advanced access to preservables, but also in the broader sense to research into e.g. Linked Open Data on the web, the Semantic Web, knowledge representation and knowledge management.

Finally, this document also **explores how all of these investigations relate to the other relevant research activities** within PERICLES and provides respective **open-source implementations and experimental results** that validate all the above.

2.2. Relation to other Work Packages

Similarly to the rest of the WP4 deliverables, the research results and recommendations in this deliverable are tightly interlinked to the following areas of RTD in other work packages:

- The underlying models for semantically representing content, context and semantic change are heavily based on the domain ontologies developed within **WP2**. Additionally, the work conducted within WP4 also feeds into the domain ontologies, revealing additional constructs and representations to be adopted by the latter.
- A significant portion of the proposed methodologies has been deployed on datasets provided by the **WP2** end-users of the project.
- Furthermore, our investigations also feed into **WP3** (LRM and Digital Ecosystem Model) and **WP5** (QA, policies and appraisal). Our work reported here connects with D5.3 with respect to quality assurance for semantics and user communities, prototypes for supporting change in technology, semantics and user communities, and semantic drift related risk assessment and appraisal. WP3, on the other hand, is meant both as a point of departure for semantic reasoning and as a point of feedback for evolving ontologies, showing that we are well anchored there.
- Finally, there is a strong linkage with the software tools and testbeds from **WP6** – more information on how the tools developed for our tasks connect to WP6 can be found in section 3.2.5 “Capturing evoving environments and content” in D6.5.

2.3. Relation to other WP4 Tasks

Besides the interconnections with other WPs, the work presented here is also linked to the other WP4 tasks as well. More specifically:

- The interpretation and reasoning activities in **T4.5** “*Contextualised content interpretation*” are based on the models and the evolving semantics representations described in **T4.4** “*Modelling Contextualised Semantics*”.

- Since work in **T4.5** also affects the models from **T4.4**, the latter also have an impact on **T4.3** "*Semantic content and use-context analysis*", looking at text and image content analytic methods from a context-dependent perspective.
- **T4.1** (PET) and **T4.2** (PET2LRM) feed into the T4.4 models for semantically representing context and use-context.

2.4. Document Structure

The structure of the rest of this document is as follows:

- **Chapter 3 "Contextualised Content Interpretation"** addresses two key targets set for WP4 in the DoW, i.e. "*Extract semantic descriptions about salient concepts and discover, through analysis of use context information, contextual usage-based content-links*", and "*Enrich concept-based semantic descriptions with background knowledge and derive higher level contextualized content interpretations through their integration*". For the first target, we present a theoretical framework adapted partly from classical mechanics, partly from quantum theory. In this framework, characteristic features of digital objects (such as words used as index terms) possess work content also known as *energy* inherent in forces in physics, but due to social influences contextual usage determines the actual values of system behaviour. Based on this framework, we are making a big step towards modelling the dynamics of semantic drifts on language based "forces". For the second target, the chapter introduces the applicability of semantic reasoning in DP-related workflows and paves the way for the proposed reasoning framework presented in a later chapter.
- **Chapter 4 "Quantum-Like Analysis for Contextual Content Interpretation"** Section 4.1 offers a brief overview of state-of-the-art quantum-inspired methods for text processing, followed by a definition of quantum likeness, contrasted by three research questions to decide if QL can be observed in our datasets. Next, we present an investigation into the quantum-like nature of semantic content related user behaviour by two case studies. This is followed by the description and implementation of a quantum-like model akin to particle-wave duality for semantic content classification, finally demonstrating the integration of two specific tools, Somoclu for drift detection and Ncpol2spda for entanglement detection. It also spans the intellectual space from context-dependence of semantic content to contextuality (non-commutativity) and entanglement, two QL symptoms showing up in information seeking behaviour, and arrives at a generalized "energetic" hypothesis underlying contextualized semantic content behaviour over time.
- **Chapter 5 "Semantic Reasoning for Contextual Content Interpretation"** reports our second line of research on contextualized content interpretation based on logical semantics and involving the use of semantic reasoning techniques. After a brief introduction to the background notions, the chapter presents the proposed PERICLES semantic interpretation framework that capitalizes on our adopted representations for contextualized content semantics. Three areas of contribution are presented: (a) ontological inference, namely, deriving implicit knowledge from asserted facts with the use of a reasoning engine; (b) rule-based reasoning, which is a more advanced reasoning approach that is based on rules; (c) contextualized reasoning on semantic drifts, which offers the capability of determining the "volatile" and conflicting concepts in an ontology model. Finally, the chapter also presents our proposed scheme for uncertainty management in contextualized content representations.
- **Chapter 6 "Conclusions & Future Work"** concludes with some final remarks and an account of potentially interesting directions for future work, with regard to each of the key topics discussed in the previous chapters.

3. Contextualised Content Interpretation

One of the key ideas underlying PERICLES is the context-dependent nature of interpreting semantic content, whichever its manifestations or levels of representation. This chapter presents an introductory overview of our proposed schemes and motivation for representing contextualized content semantics, under the scope of two core approaches briefly outlined below, and sets the scene for the following chapters providing definitions and explanations where necessary. According to the underlying basic idea for both approaches, as contexts evolve, they keep on influencing content behaviour, so that content maps or rule sets underlying reasoning become also ephemeral and context-dependent, although aging at very different rates. For reasons already mentioned in Section 2.1 and to be detailed below, we believe that such contributions are crucial to DP in general and to the future of LTDP in particular. The proposed approaches are tightly related to our work on semantic change (T4.4), offering a vehicle for identifying and exploring new aspects of evolving semantics to be used for the content mapping of collections, and semantic inference over such repositories.

3.1. Overview & Motivation

As already mentioned, this deliverable presents two parallel approaches for handling contextualized content interpretations in PERICLES: in Chapter 4 we will rely on distributional semantics for scalability and reliable statistics underlying the results, and in Chapter 5 on logical semantics for semantic reasoning. Logical semantics was introduced in WP3, whereas distributional semantics in WP4. For their interplay being explored in state-of-the-art research see Section 4.1 below. Distributional semantics depends on the immediate local context of features such as index terms on the one hand, and on their global proportions expressing topicality on the other hand, expressed by weighting schemes. Logical semantics, based on representing truth conditions via formal languages, integrate context information into the set of conditions, converting it into an integral part of the overall representation and, thus, allowing context to play a crucial part in semantic inference processes.

In this work we consider the Semantic Web, which is based on Descriptional Logics as the underlying formal representation, as the ultimate processing environment, where both kinds of context-dependent semantics contribute to content management and services as two complementary halves of the same problem solving effort. Thereby the research presented in this deliverable converges e.g. on tools like PROPheT [Mitzias et al., 2016] for ontology population and instance enrichment from Linked Open Data (LOD), and testing vector field semantics on such instances.

As stressed in the DoW, the proposed interpretation infrastructure will allow us to gain insightful contextualised views on content semantics, in order to derive content interpretations at higher-levels of abstraction that are closer to human perception and appraisal needs. Additionally, given that extremely long term DP is on the doorstep [Kazansky et al., 2016], quantum computing has just met quantum-inspired sentence processing [Zeng & Coecke, 2016], and the expected quadratic to exponential speedup it will bring is about to crossbreed with quantum machine learning [Wittek, 2014], it is fair to say that the modeling of contextualized content and user behaviour needs to adopt suitable new metaphors to catch up with novel opportunities. This will be attempted in Chapter 4, while Chapter 5 will look at the implications for semantic inference, keeping in mind that the two approaches presented here converge as exemplified by the ontological take on semantic drifts leading to contextualised semantic reasoning. Chapter 6 will sum up our considerations and outline interesting new openings for research.

3.2. Evolving Semantics & Physics as its Metaphor

To continue work in progress introduced in D4.4, we repeat our point of departure there that without advanced access to digital content, DP as an investment makes limited sense in any incarnation. Then, as human civilizations are built by means of language, semantics inherent in natural and artificial languages ferment social progress so that both individuals and communities live in shared semantic spaces in time. Our assumption is that such semantic spaces of topics constitute larger superstructures, an evolving semantic universe not unlike the physical one studied by astronomy and cosmology (Fig. 3-1). LTDP has to address the complete and durable preservation of such super-, hyper- and ultrastructures, but their exploration is just beginning and is in the phase of theory and tool development. This section will bring such considerations to the foreground.

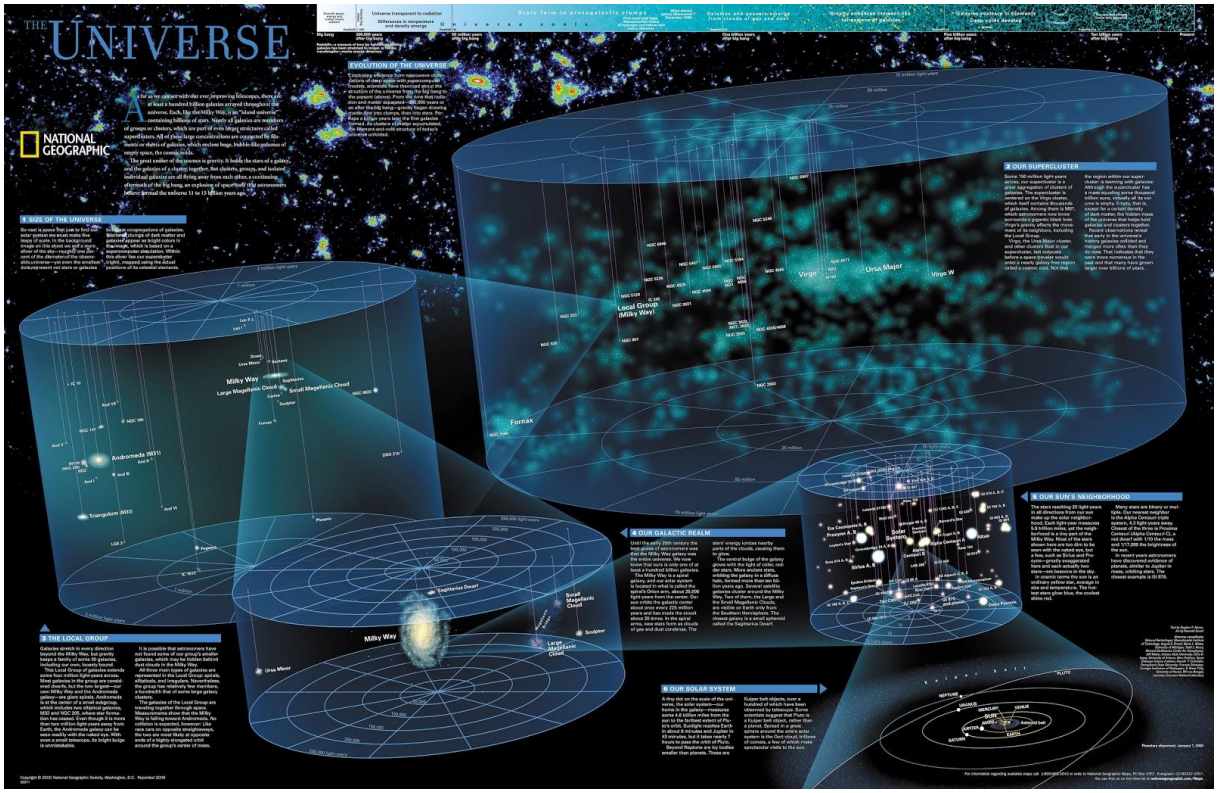

Fig. 3-1. Just like in the physical universe we experience embedded magnitudes of content that result in structures from micro- over macroscale to the ultimately largest ones, we can think of the semantic universe in similar terms. To the right bottom: the Solar System as part of the Sun's neighbourhood, itself part of the Milky Way galaxy (in the middle at the bottom), one constituent of the Local Group (to the left). The Local Group is one among many in the Local Supercluster (top right), itself again only one among many of even larger structures.

In line with the encouragement for interdisciplinary approaches by several current EU RTD calls, our results will show below that such a metaphor brings new insights. For a start, an expanding information universe as the object of study for LTPD can be described by semantics as a driving "force" only, where the dynamics of expansion is modelled on the "energy" semantic content feeds back to any such model by consecutive updates. In other words, **no other metaphor couples the experience of such an information universe with a statistically inspired methodology to model its evolution**.

To investigate the QL behavior of digital collections is part of the above strategy. By means of our metaphor, a roadmap can be outlined with semantic content related user behavior in the crosshairs,

with meaning as structuration capacity for classification, and an "energy" regime for knowledge organization. As is often the case with novel approaches, our roadmap is not consistent yet, but by the parsimonious integration of several theories over different disciplines holds benefits for the DP community.

Given a system to model semantic change, the first question is, shall the concept of interaction be applied to explain its dynamics? If yes, with interaction, the concept of force must be called in, and not just acting between any two items such as particles or wave packets, but also forming structures, and contributing to system state reconfiguration by energy. Namely with force, its work content also known as energy is as much implied as its source called the potential.

With interaction assumed between features and objects, such an impact can be measured. In Chapter 4, force will refer to action-at-a-distance type fundamental forces. These come in two kinds, gravity and electromagnetism (EM), i.e. we shall not consider weak and strong nuclear forces in our metaphor. More on them below.

Because to refer to the QL nature of digital content implies quantum mechanics (QM) as a highly successful model of how physics works, it must be mentioned here that over the past three decades, attention has been paid to the realization that the mathematics used to account for subatomic particle-wave behavior partly applies to the atomic realm and beyond too, including societies and socially grounded phenomena such as finance, economics, cognition, decision theory and language [Aerts et al., 2006; Khrennikov, 2010; Mugur-Schächter, 2014][2]. Whereas more language related applications will be mentioned in Section 4.1, we stress that most of these efforts focus on theory development and there is only a limited number of *in vivo* sightings/findings available.

Exploring physics as a metaphor to study evolving semantics inherent in language change comes with a paradox though. Apart from hunting for QL clues from early on, we have been vexed to find possible sources and explanations of the "energy" aspect implicit because QM is so strictly bound with the concept of electrons jumping between orbitals with specific energy levels underlying chemical structuration. However, this ride has been neither easy nor controversy-free. Namely our first suspect was a dipole type of a force familiar from QM, such as EM or spin – but as far as we are aware of, semantic content with such a nature is not being addressed by the trade, i.e. the weighting schemes capturing semantic content on the most elementary level record occurrence rates and normalize these between 0 and 1, but not to -1. (The only example where negative correlations pop up as a similarity measure is the Generalized Vector Spaces Model (GVSM) [Wong et al., 1985..) The importance of this is simple: with semantic similarity mapped between -1 and +1, one can model the interplay of two forces, an attractive and a repulsive one, so thereby similarity could be conceived as a "dipole" phenomenon represented by a certain distribution of positive and negative term values, standing for some balance between those forces[3]. One formula that describes such a situation in physics is Coulomb's inverse square law that describes force interacting between static electrically charged particles, in its scalar form:

$$F = k_e \frac{q_1 q_2}{r^2}$$

where k_e is Coulomb's constant, q_1 and q_2 are the signed magnitudes of the charges, and the scalar r is the distance between those charges. The force of interaction between the charges is attractive if

[2] For a current list of macroscale phenomena where QM is anticipated to be at work, see http://www.bbc.com/earth/story/20160715-organisms-might-be-quantum-machines.

[3] The utilization of the energy concept in ML goes back to the use of potentials. In the examples we are considering, there are two kinds thereof, Coulomb potential vs. gravitational potential, so that decision making (classification, categorization) is minimum or maximum seeking by gradient descent or ascent on a hypersurface, constructed from statistics describing the event space. Whereas gravitational force assumes energy from the mass of particles (i.e., documents) in a cluster, Coulomb potential presupposes the dipole nature of entities such as belonging to vs. not belonging to a class.

the charges have opposite signs (i.e. *F* is negative) and repulsive if like-signed (i.e. *F* is positive) (Fig. 3-2).

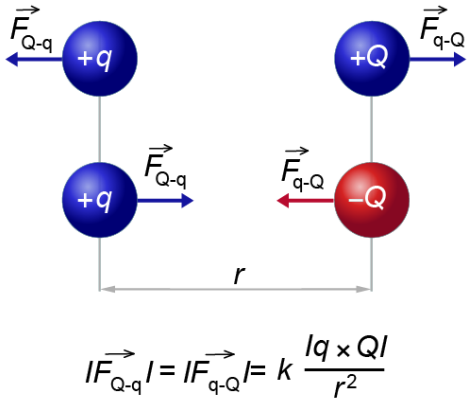

Fig. 3-2. The absolute value of the force *F* between two point charges *q* and *Q* relates to the distance between the point charges and to the simple product of their charges. The diagram shows that like charges repel each other, and opposite charges attract each other.

With no such positive-negative value distribution to characterize term behavior toward one another, we defaulted on the testing of Newton's universal law of gravitation as part of Newtonian dynamics, another inverse square law of a very similar form, but simulating a "monopole", i.e. attractive-only force:

$$F = G\frac{m_1 m_2}{r^2}$$

Here, *F* is the force between the masses; *G* is the gravitational constant, m_1 is the first mass, m_2 is the second mass, and *r* is the distance between the centers of the masses (Fig. 3-3).

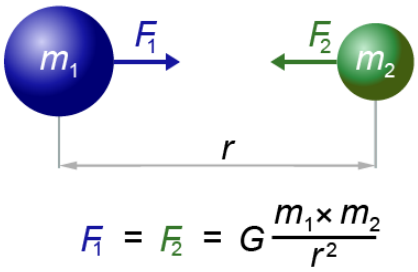

Fig. 3-3. Diagram of two masses attracting each other.

As we will see below, the findings of this approach (in Experiment 3) are very interesting, but this is also where controversy creeps in. Namely in modern physics, QM and the General Theory of Relativity (GTR) – the latter building its world view on gravitation – are still contestants for a unified theory of physical reality, with efforts such as Quantum Field Theory (QFT) trying to bridge the gap between their conflicting concepts by turning to the Special Theory of Relativity. The fact that in Experiments 1-2 we can offer evidence for the QL nature of the results repeats in a loose sense this deep contradiction. With Experiment 4 possibly showing an overlap between this QL nature and semantic drifts modelled on gravitation, the mystery deepens.

The GTR as part of our conceptual framework is not a must because QM would be imbalanced without it, nor because we could exploit the relativistic equivalence between mass and energy. There is nothing relativistic about language and language-based communication, including semantics. On the other hand, there is apparently a link between semantic (Mercer) kernels as similarity measures,

their capacity to create a separating hyperplane between classes of objects in Support Vector Machines (SVM) [Vapnik, 1998; Schölkopf & Schmola, 2002], the curvature of these planes, and the Riemann metric tensor from GTR whose job is to describe the curvature of space according to the mass bending it [Moschitti, 2010; Amari & Wu, 1999; Eklund, 2016; Williams et al., 2005].

This means that, being short on the functional equivalents of EM or spin in semantics, the "energetic" aspect of QL symptoms of digital content is not explained for the time being, whereas at the same time the "gravitational" approach indicates the presence of a strong external potential, i.e. contextual dependency of the semantic drift on its social embeddings. Very likely we are observing different types of social forces at work, i.e. "energy" as a driver of evolving semantics is there but it is not the one underpinning our QL findings. This situation is conceptually framed by the idea of social mechanics, a paradigm increasingly used to model social forces and important here to add contextual dependence to our model of dynamic semantics.

We stress that it was not our aim to reconcile QM and GTR in a new conceptual frame, rather what we had in focus has been the semantic drift and its dynamics. Therefore what we shall report as QL results below are such because of some characteristic symptoms measured, not because of the nature of the "forces" or interactions at work in the digital collections experimented with. Still, it will add to their QL appeal that QM interpretations which claim to address both gravity and EM, including particle-wave duality, seem to be applicable to digital content (e.g. Bohmian mechanics and its variants), but without the energy levels typical of QM.

With the above as background, we shall seek answers to three related research questions:

- Q1: Do any of the selected QM symptoms show up in digital collections potentially significant for evolving semantics and digital preservation? Does their presence justify the QL label?
- Q2: Do any of the selected GTR/CM symptoms show up in digital collections potentially significant for evolving semantics and digital preservation?
- Q3: Is there an overlap between the QL and GTR/CM nature of digital collections potentially significant for evolving semantics and digital preservation?

A list of criteria, typical for QM, will define QL with regard to ES for Chapter 4 in Section 4.2. The section plan for Chapter 4 where we answer the above is as follows. We begin with the definition of QL and examples of QM and computational linguistics to model sentence meaning and semantic reasoning. Then we report four experiments to prove our point, based on eye gaze test data, citations and the Tate datasets by Somoclu [Wittek et al., 2013] and Ncpol2spda [Wittek, 2015] as exploration tools:

- **Experiment 1**: one aspect of QL, non-commutativity, on eye gaze (by Ncpol2spda);
- **Experiment 2**: another aspect of QL, entanglement without nonlocality on citations (by Ncpol2spda);
- **Experiment 3**: continuing D4.4 (drifts), gravitation leading to external potential (EP) and "particle-wave duality" as another aspect of QL (by Somoclu);
- **Experiment 4**: combining Experiments 2-3, entanglement without nonlocality over drifts (by Somoclu and Ncpol2spda).

We shall offer evidence for the QL nature of digital content inasmuch as it does not behave completely as described by QM because there is only partial overlap between semantic and QM criteria. But neither does it fully comply with CM, because it uses a non-constant, relative mass concept instead of specific, constant values to characterize semantic features. As a consequence of this, one can augment de Saussure's concept of a linguistic sign as the unity of form and content by a third component, its structuration capacity. This capacity will denote the energy a.k.a. work content inherent in a semantic unit (such as an index term or a machine learning feature), representing amounts of content investment for the reconfiguration of semantic spaces during update.

3.3. Semantic Reasoning for DP

DP's primary aim is to secure long term access to cultural heritage content by ensuring that **the content's significant characteristics are not lost over time** [Engen et al., 2015]. However, rather than plainly protecting and ensuring archiving and accessibility, DP also requires **constant enrichment of the content with explicit and implicit semantic associations** from within the global data collections. Moreover, DP workflows and activities depend upon certain vital processes, such as **semantic drift**, **risk assessment**, **descision support** and **quality assurance systems**.

On the other hand, the Semantic Web offers an arsenal of powerful mechanisms towards the discovery, amplification and semantic interconnection of knowledge. Semantic technologies have also been noted as essential enablers for reasoning of actionable knowledge from multiple heterogeneous information sources and disparate domains, and foster interoperability amongst a variety of applications and systems [Maarala et al., 2016]. In this direction, the semantic reasoning mechanisms excel at inferencing logical consequences from formally represented knowledge (mainly ontologies) and successfully back the variety of DP prerequisite processes, as described in the previous paragraph. Therefore, we consider the Semantic Web as the ultimate knowledge repository and reasoning arena to be used for the evolving DP processes.

3.4. Chapter Summary

We outlined the work conducted within task T4.5 of WP4, with proposed approaches for contextualised content interpretation targeting insightful contextualised views on content semantics. This is achieved through the adoption of appropriate context-aware semantic models developed within the project, and via enriching the semantic descriptions with background knowledge, deriving thus higher level contextualised content interpretations that are closer to human perception and appraisal needs. All these contributions are tightly coupled with the other PERICLES work packages for takeup and implementation. In the following chapters, first we present our findings about evolving semantic content modelled by physical concepts, then about advanced semantic reasoning. Both chapters rely on earlier work on the LRM and semantic drifts. This continuum of linked solutions helps us to spell out future research directions as a sustainability effort.

4. Quantum-Like Analysis for Contextual Content Interpretation

Quantum mechanics, quantum physics, and quantum theory are often used as synonyms, but there are subtleties worth mentioning. In our reading, quantum mechanics is the most specific field that focuses on non-relativistic particles described by the Schrödinger equation whose dynamics are unitary. Quantum physics is broader and includes, for instance, quantum optics, where unitary dynamics are approximated by linear and nonlinear optical systems. Quantum theory is the broadest -- we view it as the mathematical framework without necessary references to any physical system. Quantum-likeness borrows ideas and metaphors from quantum theory that fit a certain task. We agree that our take on the relationship of these fields is subjective and we do not argue its correctness: we included this clarification to avoid misunderstandings.

Our major target in this line of research has been the investigation of the QL nature of digital objects and the development of context-dependent, quantum-based models for semantic content classification. To this end, we shall depart from content dynamics typical of evolving semantics, exemplified on semantic drifts, and model them on force fields with energy content, matching the concept of lexical fields in linguistics with that of vector fields as used in physics. Above in Section 3.4, we have outlined our theoretical considerations for a framework for hypothesis testing. In Chapter 4, we look into drift dynamics from a part classical (gravitational), part non-classical (Quantum Theory-inspired) angle, focusing on the following:

1. A proof-of-concept experiment uses classical mechanics (CM) and gravitation, to show how the concept of term “energy” can be applied to indexing features. The resulting semantic potential expresses the “first among equals” principle, a social phenomenon historically well documented, i.e. points out the most important features and documents among similar ones;
2. Another proof-of-concept experiment identifies two key symptoms of Quantum Theory (QT) during information seeking, the uncertainty relation leading to non-commutativity (contextuality) in eye gaze fixation, and entanglement as non-classical correlation in citation patterns. Thereby, QL usage-specific information seeking behaviour becomes the embedding context for the above semantic potential;
3. These findings, i.e. a combination of energetic indexing feature behaviour modelled on gravity, combined with an evolving QL usage context modelled on QT, invite particle-wave duality as a parallel for the interpretation of evolving semantics. Using several examples from Ehrenfest’s theorem to the Schrödinger equation, we suggest Bohmian mechanics as an applicable theoretical frame for continuing work;
4. As the concepts of force and field are intimately linked with the energy content of a system state called its Hamiltonian, first we show that the gravitational methodology is able to identify external forces acting upon system states, and then we explain how the Hamiltonian plays an important role in a QT framework, prominently in particle-wave duality. Based on this, we conjecture that an “energy regime” could account for knowledge dynamics in general, doing the bookkeeping of feature investment vs. structuration results. We shall round off this conjecture by comparing evolving constellations of semantic content to those of physical content observed by astronomy.

This means that in accord with T4.5.1 in the DoW and its main lines of inquiry, we continue to address the dynamics of evolving semantics. Our goal is to design a method by which we can measure conceptual investment into the configuration and reconfiguration of semantic spaces, and thereby identify new preservables for LTDP. Central to scalable knowledge organization, we shall refer to this recurrent activity as *structuration*.

In nature, both structuration and its dynamics – including phenomena described by QM – depend on *energy*. To find out if patterns in our data are similar to those observed in physics in general, and QM in particular, is part of the investigation. This will be our compass when we next briefly survey relevant research directions. We start with a survey into the state-of-the-art in quantum-like methods for text interpretation.

4.1. Quantum-like Methods for Text Interpretation

Apart from sporadic attempts to link word senses with QM and treat them as quantum states [Bruza and Woods, 2008; Wittek & Darányi, 2011; Blacoe et al., 2013], most of the related research is focusing on sentence semantics, prominently by compositional semantics. Sentence semantics is currently not a commercial approach in IR or ML, but phrase-based indexing is coming up [Mikolov et al., 2013], and as it could be exploited for LRM-based statements, a brief detour will be useful. It holds for practically all approaches listed below that they employ ideas from QM – or quantum theory (QT), the mathematical foundation upon which of QM is built but devoid of its physical content – to encode the formal side of language, and its semantics only thereby.

Compositional semantics relies on the principle of compositionality. In mathematics, semantics, and philosophy of language, this principle argues that the meaning of a complex expression is determined by the meanings of its constituent expressions and the rules used to combine them, i.e. its structure [Szabó, 2013]. From this principle, ascribed to Frege, the general direction of research is toward **compositional distributional semantics** (**CDS**), combining philosophy, quantum mechanics, information theory and cognitive linguistics. As the name implies, CDS tries to construct a sentence meaning representation based on the distributional hypothesis used to encode word meaning in vector space.

One of the counterintuitive aspects of quantum mechanics is that it is *non-local.* That is, particles can influence each other's behaviour at distances from which this should not be possible according to common sense [Einstein et al., 1935]. Such *entanglement* can be seen as a channel for passing information from one particle to another. Indeed, quantum computation and quantum information theory are based on this premise. A basic unit of information is in fact defined there, dubbed a *qubit,* and the change in its information content under various operations is studied, even going so far as to quantify how much information a channel can transmit [Nielsen & Chuang, 2000]. On the other hand, there are two major approaches to the formal analysis of natural language: one of a **logical** nature (Dowty et al., 1981], and one of a **distributional** kind relying on *vector spaces* as models of meaning [Schütze, 1998; Clark & Pulman, 2007]. These two schemes have complementary features. The logical model is “compositional”: the meaning of a sentence is a function of the meanings of its words-but says nothing about the lexical meaning, i.e. the meanings of individual words. The vector space model constructs meanings of individual words by counting co-occurrence with words used often in a given corpus of text, but does not address meanings of strings of words. Using inspiration from quantum teleportation [Clark et al., 2013] and the concept of pregroups from logic [Lambek, 2011], various authors work on composing sentence encodings from word vectors by tensor algebra [Coecke et al., 2010; Coecke et al., 2013; Grefenstette, 2013; Blacoe 2015b]. In this process, the concepts of interaction and non-locality are central to underpin the QL claim. To obtain the meaning of a simple sentence, e.g. “John likes Mary”, the verb first has to interact with its subject and object, so that this interaction results in a grammatical structure. But in a negative sentence, e.g. “John does not like Mary”, the subject does not immediately precede the verb, as it would in the positive case. Instead, they are separated by 'does' and 'not'. The compact structure morphisms reconnect the two

via a *non-local correlation* in such a way that the verb can still act on its subject from a distance, and the application of 'not' is delayed [Heunen et al., 2013][4].

Parallel to the quest to find vector-based phrase and sentence encodings as well [Mikolov et al., 2013], there are ongoing efforts to combine distributional and compositional semantics [Clark & Pulman, 2007; Blacoe & Lapata, 2012]. For instance Sadrzadeh and Grefenstette propose a hybrid approach based on categorical methods also applied to the analysis of information flow in quantum protocols. The mathematical setting stipulates that the meaning of a sentence is a linear function of the tensor products of the meanings of its words. The applicability of these methods is demonstrated via a toy vector space as well as real data from the British National Corpus and two disambiguation experiments [Sadrzadeh & Grefenstette, 2011]. Fyshe et al. explore the utility of combining topical information (e.g., documents in which a word is present) with syntactic/semantic types of words (e.g., dependency parse links of a word in sentences) to compose adjective-noun phrases [Fyshe et al., 2013]. Blacoe introduces a tensor-based model that constructs compositional representations for sentences of arbitrary length. Representations for individual words are captured by distributions of dependency neighborhoods which encode sufficient lexical and structural information to perform semantic composition [Blacoe, 2015b].

We note in passing that before tensors, circular convolution was applied to sentence encoding [Plate, 1991]. By its complex valued vector representation, the notions of probability, logic and geometry are integrated within a single Hilbert space representation by [De Vine & Bruza, 2010]. This has inspired Cohen and Widdows to come up with Predication-based Semantic Indexing (PSI), an approach to generating high-dimensional vector representations of concept-relation-concept triplets. In this paper, we develop a variant of PSI that accommodates estimation of the probability of encountering a particular predication (such as 'uoxetine TREATS major depressive disorder') in a collection of predications concerning a concept of interest (such as a major depressive disorder). PSI leverages reversible vector transformations provided by representational approaches known as Vector Symbolic Architectures (VSA). To embed probabilities they develop a novel VSA variant, Hermitian Holographic Reduced Representations, with improvements in predictive modeling experiments. The probabilistic interpretation this facilitates reveals previously unrecognized connections between PSI and quantum theory - perhaps most notably that PSI's estimation of relatedness across multiple reasoning pathways corresponds to the estimation of the probability of traversing indistinguishable pathways in accordance with the rules of quantum probability [Cohen & Widdows, 2015]. This heralds a combination of analogical reasoning over triples with quantum theory already available for the processing of medical literature.

A specific direction to model lexical semantics on QT is in [Blacoe et al., 2013]. They explore the potential of quantum theory as a formal framework for capturing lexical meaning by creating a novel semantic space model that is syntactically aware, takes word order into account, and features key quantum aspects such as superposition and entanglement by probabilities. They also define a dependency-based Hilbert space and show how to represent the meaning of words by density matrices (density operators) that encode dependency neighborhoods. This approach is extended to an unsupervised model that learns word meanings from a large corpus of dependency-parsed English sentences in an unsupervised way. The model is able to detect whether a pair of sentences constitutes a paraphrase, and if there is entanglement among syntactic relations within linguistic density operators [Blacoe, 2015a].

In the next section we continue our investigation into the quantum-like nature of digital objects and development of quantum-based models for semantic content classification. Specifically, we shall

[4] Once vectors for the meanings of sentences are built, one can use the cosine similarity measure to automatically derive synonymous sentences, a task which has applications in automating language-related tasks such as translation and paraphrasing.

focus on the interaction of content and information seeking behaviour as its “outer layer” of evolving context.

4.2. Investigation into the Quantum-like Nature of Semantic Content Related User Behaviour

As for the quantum-inspired representation of sentence meaning there are strong indications in the literature, these encourage us to look for more evidence of applicability below sentence level but being dependent on context. We shall do so below, because "*techniques used in quantum interaction often involve modelling concepts in vector spaces, and finding ways to model concept combinations as operations on vectors is crucial when trying to apply quantum-like techniques to increasingly challenging and sophisticated information modelling tasks*"[5].

Keeping in mind that [Blacoe et al., 2013]’s solution treats word senses as kets and corresponding density matrices (density operators), and disambiguates them based on dependency grammar in a tree bank, hence their word senses are distributional and occurrence rate related, below we shall tap into the very same mass-like source of “energy” distinguishing word senses by a different approach. Our quest will cover the “energetic” aspect of word meaning and semantic content related information seeking behavior, respectively.

For this deliverable, we define QL as follows: “The applicability of concepts from quantum theory and their mathematical formalism to non-QM data of social origin, including DOs, their metadata, content patterns and their evolution, and user behaviour”. A partial compliance with a list of criteria below, typical for QM, will define QL with regard to ES:

- The uncertainty relation (cf. contextuality, non-commutativity) should be present in the data;
- Entanglement should show up;
- The concept of particle-wave duality should be applicable.

We shall report two case studies of information seeking behaviour, one on eye gaze fixation data, the other on temporal correlations in citation patterns.

4.2.1. Eye Gaze Fixation during Information Searching & Contextuality

Below we review an article co-authored with non-PERICLES partners [Wittek et al., 2016a]. Its importance lies in the fact that it identifies *noncommutativity* (*contextuality*) in information seeking behaviour, finding that the outcome of a search strategy is dependent on the sequence of its steps. Such noncommutativity is a typical hallmark of QM. Finally, information seeking is a combination of semantic content and evolving user behaviour, not studied as a mix in D4.4 but interesting for DP in its own right, therefore addressed here.

Information foraging connects optimal foraging theory in ecology with how humans search for information. The theory suggests that, following an information scent, the information seeker must optimize the tradeoff between exploration and exploitation of information items by repeated steps in the search space vs. exploitation, using the resources encountered. We conjecture that this tradeoff characterizes how a user deals with uncertainty and its two aspects, risk and ambiguity in economic theory. Risk is related to the perceived quality of the actually visited patch of information, and can be reduced by exploiting and understanding the patch to a better extent. Ambiguity, on the other hand, is the opportunity cost of having higher quality patches elsewhere in the search space. The aforementioned tradeoff depends on many attributes, including traits of the user: at the two

[5] Dominic Widdows (Grab Technologies, US), private communication.

extreme ends of the spectrum, analytic and holistic searchers employ entirely different strategies. The former type focuses on exploitation first, interspersed with bouts of exploration, whereas the latter type prefers to explore the search space first and consume later. Based on an eye-tracking study of experts' interactions with novel search interfaces in the biomedical domain, we demonstrate that perceived risk shifts the balance between exploration and exploitation in either type of users, tilting it against vs. in favour of ambiguity minimization. Since the pattern of behaviour in information foraging is quintessentially sequential, risk and ambiguity minimization cannot happen simultaneously, leading to a fundamental limit on how good such a tradeoff can be. This in turn connects information seeking with the emergent field of quantum decision theory.

Introduction

Searching for food is a common pattern of behaviour: humans and animals share dedicated cognitive mechanisms to find resources in the environment. Such resources are distributed in spatially localized patches where the task is to maximize one's intake, that is, knowing when to exploit a local patch versus when is it time to move on and explore one's broader surroundings.

In humans, the underlying neuropsychological mechanisms result in cognitive searches, such as recalling words from memory [Hills et al., 2012; Hills et al., 2015]. As part of users' information seeking behaviour, the concept of information foraging describes the above quest by a similar strategy [Pirolli & Card, 1999].

Key to the understanding of decisions by a consumer of information is that they are subject to uncertainty: his or her knowledge of the environment is incomplete, so the resulting decisions must go back to perceptions and certain heuristics. By turning to classical works in economy, we can identify two facets of this uncertainty, namely risk and ambiguity [Knight, 1921; Ellsberg, 1961]. Their interpretation according to the foraging scenario is in place here.

Briefly, *risk* would be the quality of the current patch and our fragmented perception of it. Is the place of good quality? Should one stay here or move on? Since we are already at the preselected location, we do have prior information about it. A risk-minimizing behaviour will favour exploitation over exploration, staying longer at individual locations, potentially losing out if outstanding patches remain unvisited. *Ambiguity*, on the other hand, is related to opportunity cost, the price of not foraging elsewhere. "Elsewhere" refers to the rest of the unknown distribution which is not observed at the moment. A human forager who wants to reduce ambiguity first will jump around different patches and explore more, learning as much as possible about the information distribution while reducing the associated uncertainty. This behaviour will not stop at the first good enough patch.

Resonating with the aforementioned, our working hypothesis below will be that if animal foraging is subject to uncertainty, and information seeking is an essentially identical activity in a different context, then a limit to simultaneous risk and ambiguity minimization must apply to information foraging as well. This limit emerges from the sequential and incompatible nature of the decisions made to minimize these two aspects of uncertainty. We will demonstrate our point on eye tracking data in study of user interactions with novel search interfaces for biomedical information search. The incompatible decisions are similar to noncommuting measurements in quantum mechanics where they give rise to the uncertainty principle; thus our work connects information foraging and information seeking behaviour to the thriving field of quantum decision theory [Yukalov & Sornette, 2008; Bruza et al., 2009; Khrennikov, 2010; Busemeyer & Bruza, 2012; Ashtiani & Azgomi, 2015].

The Origins and Application Areas of Uncertainty

A decision in the presence of uncertainty means that the outcome cannot be fully predicted before the decision is made. Multiple possible outcomes can occur, and our knowledge of the probability distribution only allows for a limited characterization of uncertainty. Following the work by [Knight, 1921; Ellsberg, 1961; Camerer & Weber, 1992], we can distinguish between two fundamental aspects

of uncertainty, aforementioned ambiguity and risk. The simple definition of risk is uncertainty with known probabilities, a certain a priori probability for a given outcome. Ambiguity is also probabilistic but less well defined, generally associated with events that the decision maker has even less information about than the risk of outcomes. The two aspects are also called expected and unexpected uncertainty. Dealing with unexpected uncertainty involves a more subjective evaluation of probabilities. In the case of ambiguity, less information is available, and expected utility is harder to estimate. Not knowing crucial information, such as the probability distribution of the outcomes, is a frightening prospect which explains why most people are ambiguity-averse [Ellsberg, 1961]. The two forms of uncertainty are so different that dealing with risk and ambiguity are supported by distinct neural mechanisms in humans [Huettel et al., 2006].

Apart from this probabilistic nature of decisions in an uncertain environment, there is an even deeper form of uncertainty: the kind we normally refer to in the context of quantum mechanics (QM). Some measurements on a quantum system are simply incompatible: measuring one aspect of the system prevents us from learning more about another aspect thereof, explored by a different measurement.

As stated by [Folland & Sitaram, 1997] in what constitutes the basis of this brief overview, "*There are various mathematical aspects of the uncertainty principle, including Heisenberg's inequality and its variants, local uncertainty inequalities, logarithmic uncertainty inequalities, results relating to Wigner distributions, qualitative uncertainty principles, theorems on approximate concentration, and decompositions of phase space*". It is partly a description of a characteristic feature of quantum mechanical systems, partly a statement about the limitations of one's ability to perform measurements on a system without disturbing it, and partly a meta-theorem in harmonic analysis that can be summed up as follows: "*A nonzero function and its Fourier transform cannot both be sharply localized*". Therefore the principle leads to mathematical formulations of the physical ideas first developed in Heisenberg's seminal paper of 1927 [Heisenberg, 1927], explored from many angles afterwards.

Incompatible measurements mean that certain observations on a system do not commute: by making an observation, we are making a second one in the context created by the first. In other words, incompatibility, noncommutativity, and contextuality are closely related concepts.

Noncommutativity allows the definition of an alternative event algebra or logic, which in turn leads to applications in decision theory [Bruza et al., 2009; Busemeyer & Bruza, 2012]. This line of research is part of a broader trend of applying the mathematical framework of quantum mechanics in domains outside physics [Khrennikov, 2010].

Uncertainty and Foraging Decisions

We are especially interested in how risk and ambiguity appear in sequential decisions. Simultaneous or coordinated decision making, on the other hand, is more complex, being less common among animals because it involves comparative evaluation. Pointing at a major difference between the animal kingdom vs. man [Kolling et al., 2012] showed that humans are able to choose between these two models in uncertain environments. A foraging scenario is a good example of sequential decision making: food resources are available in patches, and a forager must find an optimal strategy to consume the resources. There is a cost associated with switching from one patch to another. Uncertainty relates to the quality of the current patch, the quality of background options -- the opportunity cost of not foraging elsewhere -- and the environment is also subject to changes. The forager has to minimize the tradeoff between exploitation of a patch versus exploration of background options. The pattern is not restricted to food consumption: for instance, it pertains to mate selection, retrieving memories, and consumer decisions. In fact, the same neural mechanism can serve these different functions [Adams et al., 2012].

Optimal foraging theory gives the strategy to follow if the probabilities can be estimated and updated by the forager [McArthur & Pianka, 1966; Charnov, 1976]. Ambiguity alters the behaviour: for example, unexpected forms of uncertainty may trigger more exploration [Cohen et al., 2007]. We would like to see how ambiguity and risk can be minimized in sequential decisions, and how that affects exploration and exploitation.

Many decisions require an exploration of alternatives before committing to one and exploiting the consequences thereof. This is known as foraging in animals that face an environment in which food resources are available in patches: the forager explores the environment looking for high-quality patches, eventually exploiting a few of them only. The decisions take place in an uncertain environment: ambiguity about the quality of patches and the risk of not foraging at better patches force the forager to accept a tradeoff.

Risk-sensitive foraging is not exclusive to animals, human subjects also show similar behavioural patterns [Pietras et al., 2003; Rushworth et al., 2012]. An optimal solution between exploration and exploitation is generally not known, except in cases with strong assumptions about both the environment and the decision maker [Cohen et al., 2007]. The tradeoff between exploration and exploitation is also known as the partial-feedback paradigm, linking the decision model to the description-experience gap [Hertwig & Erev, 2009]: people perceive the risk of a rare event differently if the probability distribution is known (decision from description) vs. when they have to rely on more uncertain information (decision from experience).

Information Seeking as a Form of Foraging

To take the next step in our working hypothesis, below we shall look at a scenario where seeking was exercised by gaze fixation at segments of user interfaces with significant elements of content, and show that underlying the seemingly random walks of eye gaze on the screen, there is order in the patterned data inasmuch as a certain typology of user behaviour applies to them.

The information foraging nature of the data was recognized by eye tracking analysis, based on the concept of information scent, operationalized as “the proportion of participants who correctly identified the location of the task answer from looking at upper branches in the tree" in a study of user interactions with visualization of large tree structures [Pirolli et al., 2000]. [Pirolli et al., 2001] provided further theoretical accounts for scanpaths from cognitive perspectives in which users were able to find information more quickly when strong information scent was detected. [Chi et al., 2001] built a computational model for user information needs and search behaviour based on information scent, and the model and algorithm were evaluated by simulated studies. More recently, the modeling of user search behavior using eye tracking techniques has focused on levels of domain knowledge, user interests, types of search task and relevance judgments in search processes [Cole et al., 2010; Cole et al., 2013; Gwidzka, 2014; Vakkari et al., 2014; Zhang et al., 2015]. However, there is still limited understanding of the effect of individual differences and user perceptions of search tasks on eye gaze patterns in information search. [White, 2016a; White, 2016b] provided a review of information foraging and user interactions with search systems.

The eye gaze patterns, an indicator of user attention and cognitive processes have been extensively studied for designing user interfaces, such as the functional grouping of interface menu [Brumby & Zhuang, 2015; Goldberg & Kotval, 1999], faceted search interface [Kemman et al., 2013; Kules et al., 2009] and comparison of interface layouts [Kammerer & Gerjects, 2012]. Information retrieval researchers have been concerned with users' attention to the ranking position of documents and different components of search engine results page (SERP) [Cutrell & Guan, 2007; Dumais et al., 2010; Kim et al., 2016; Lorigo et al., 2008; Savenkov et al., 2011]. These studies generally suggest that there is no significant difference in users' eye gaze patterns on comparisons of search interface layouts, and users' attention to elements of interfaces depends on the length and quality of snippets on SERPs, as well as the displayed position of search results.

User Experiment

We designed a study to investigate user gaze and search behaviour in biomedical search tasks, with particular reference to the user's attention to and use of the document surrogates (i.e., Medical Subject Headings (MeSH) terms, title, authors, and abstract). A total of 32 biomedical experts participated in the controlled user experiment, performing searches on clinical information for patients. The participants were mostly students with search engine experience and some academic background in the biomedical domain. We used a 4 x 4 x 2 factorial design with four search interfaces, controlled search topic pairs and cognitive styles. A 4 x 4 Graeco-Latin square design was used [Fisher, 1935] to arrange the experimental conditions. Each user was assigned 8 topics in total, with a 7-minute limit for each topic, and the experiment took about 90 minutes in total.

Search Interfaces

Participants searched on four different search interfaces, with a single search system behind the scenes. The four search interfaces were distinguished by whether MeSH terms were presented and how the displayed MeSH terms were generated:

Interface “A” mimicked web search and other search systems with no controlled vocabulary. This interface had a brief task description at top; a conventional search box and button; and each result was represented with its title, authors, publication details, and abstract where available. Full text was not available, so the results were not clickable. Users judged their success on the titles and abstracts alone.

Interface “B” added MeSH terms to the interface. After the user's query was run, MeSH terms from all results were collated; the ten most frequent were displayed at the top of the screen. This mimics the per-query suggestions produced by systems like ProQuest[6]. MeSH terms were introduced with “Try:” and were clickable: if a user clicked a term, his or her query was refined to include the MeSH term and then re-run. It was hoped that the label, and the fact they work as links, would encourage users to interact with them.

Interface “C” used the same MeSH terms as “B” but displayed them alongside each document, where they may have been more (or less) visible. It is a hybrid of interfaces “B” and “D”.

Interface “D” mimicked EBSCOhost[7] and similar systems that provide indexing terms alongside each document. As well as the standard elements from interface “A”, interface “D” displayed the MeSH terms associated with each document, as part of that document's surrogate. Again, terms were introduced with “Try:” and were clickable.

Each interface was labelled with a simple figure: a square, circle, diamond, or triangle, which was referred to in the exit questionnaire. A save icon alongside each retrieved document was provided to collect user perceived relevant documents.

Search Topics

Search topics used here were a subset of the clinical topics from OHSUMED [Hersh et al., 1994], originally created for information retrieval system evaluation. The topics were slightly rewritten so they read as instructions to the participants. Topics were selected to cover a range of difficulties.

Procedure

Participants were given brief instructions about the search task and system features, followed by a practice topic and then the searches themselves. They were informed that the test collection is incomplete and out-of-date since the OHSUMED test collection [Hersh et al., 1994] was used, with

[6] For example, see http://www.proquest.co.uk/en-UK/products/brands/pl_pq.shtml
[7] http://www.ebscohost.com/

MEDLINE data from 1987 to 1991. User interaction data recorded included: all queries, mouse clicks, retrieved and saved documents, time spent, and eye movements. Electroencephalogram (EEG) readings were also captured.

Background and exit questionnaires collected demographic information and asked participants about their perception of the search process. Participants' opinions of the tasks and the interfaces were sought. Finally, information on participants' cognitive styles was collected by a computerised test [Peterson et al., 2003; Peterson, 2005], which took a further 15 minutes to complete.

Results of Search Behaviour and Eye Gaze

Overall, results from the above experiment supported the hypothesis that search interfaces have significant effects on eye gaze behaviour in terms of proportion of fixations in reading time. For a detailed overview, please visit Sections 4.1-4.4 in our publication[8]. Table 4-1 displays the connection between search behaviour and eye gaze fixation.

Table 4-1. The connection between search behaviour and eye gaze fixation.

	# of queries issued	# of MeSH queries issued	# of mouse clicks	# of pages viewed	# of documents saved
Title	—	—	❍	❍	—
Author	—	—	—	—	—
Abstract	—	❍	—	—	●
MeSH	●	—	—	—	❍

Note. The relationship is not statistically significant (—), positively significant (●), or negatively significant (❍).

Their main findings in this data sample were as follows:

- When users perceived their search tasks as difficult, they did not attend to all content elements in documents;
- Searchers with different cognitive styles may use different search strategies in an environment with uncertainty they perceive as difficult;
- Search behaviour associated with expanding mental efforts like issuing MeSH terms and viewing SERPs has not changed according to the uncertainty within the environment, such as perceived search task difficulty;
- Certain search behaviour types, such as issuing queries and MeSH terms that involve notable mental efforts and exploitation of resources, are correlated with changes in eye gaze patterns.

These findings indicate distinct strategies in dealing with uncertainty, possibly changing from exploration to exploitation and vice versa, and therefore corroborate our hypothesis that the corresponding observations do not commute, which, in turn, means that information foraging is a form of quantum-like behavior. These two statements will be substantiated in the next two sections.

Different Strategies and Noncommuting Observations

In the above eye tracking study, the document surrogates and the four layouts characterize different perceptions of risk of information patches, gazing time being a good figure of merit for exploitation. Exploration is the jumping gaze combined with a repeated query as these reduce overall ambiguity. There is evidence that holistic users prefer to get an overview of tasks before drilling down to detail, whereas analytic users look for specific information. These two extreme user behaviours rely on the two measurement operators, namely risk- vs. ambiguity reduction, in different order, proving

[8] https://arxiv.org/abs/1606.08157

noncommutativity. Unfortunately, at this point there is no significant relationship yet between the users' cognitive style and the AOIs in the study.

However, if we also change the perceived risk by varying the search interface, the picture changes. The effect of cognitive styles, interfaces and their interactions on the AOI of MeSH terms (excluding Interface A) is statistically significant in terms of cognitive style and interface interactions, and weakly significant in terms of cognitive style ($F(1,188) = 2.79$, $p < .01$). Interfaces make a statistically significant difference for the holistic style ($F(2, 111) = 6.58$, $p < .001$), and cognitive styles make a statistically significant difference in Interface B ($F(1, 62) = 5.11$, $p < .05$). The results indicate that holistic users' attention to the MeSH terms is more affected by search interfaces than that of analytic users, and this interaction effect is significant when interacting with Interface B. Thus noncommutative measurements emerge.

Information Seeking is Quantum-Like

To sum up, uncertainty as a composite of risk and ambiguity drives information seeking behaviour in a complex way, with successive decisions attempting to minimize both components at the same time. However, to find their joint optimum is not possible, because risk-prone and ambiguity-prone behaviour manifest two versions of foraging attitude, called the “consume first and worry later” (exploitation) vs. the “worry first and consume later” (exploration) types. Whichever option taken, it becomes the context of the opposite alternative, so that ambiguity minimization dependent on risk minimization vs. risk minimization dependent on ambiguity minimization yield different sets of retrieved items, i.e. the outcome of information seeking as a process is non-commutative.

For every case where this joint optimum seeking mentality influences the results, plus the decision making process that has led to a particular outcome must be preserved for future reconstruction, our findings are relevant. However, there is more to the implications of the above.

From our experiment, we have seen that two types of information seeking behaviour emerged from interaction between the cognitive apparatus and the phenomenon observed, i.e. information. This is reminiscent of the Copenhagen interpretation of QM, where interaction between the measurement apparatus and the observable cannot be reduced to zero, and the measured value is a result of (or is not independent from) interaction; again in the words of [Folland & Sitaram, 1997], “the values of a pair of canonically conjugate observables such as position and momentum cannot both be precisely determined in any quantum state.” Further, we have found that the above two types of behaviour go back to the application of two operators, risk- and ambiguity-aversion, so that by applying now this, then the other first, their sequential application leads to different results, called non-commutativity.

Moreover, as much as risk and ambiguity are two sides of the same coin, non-commutativity is an essential feature of the uncertainty principle core to quantum mechanics. Given this, our current finding hints at something potentially fundamental about the nature of browsing. At the same time, since [Dominich, 2001] proposed to treat precision and recall as complementary operators regulating the surface of effectiveness in information retrieval, whereas [van Rijsbergen, 2004] argued that relevance is an operator on Hilbert space and as such is part of the quantum measurement process, neither was our insight totally unexpected. Rather, connected to the uncertainty principle, we see noncommuting measurements to surface also in information seeking as another link to quantum decision theory [Wittek et al., 2013a; Ashtiani & Azgomi, 2014; Aerts & Sozzo, 2016].

4.2.2. Citation Behavior and Entanglement

As Experiment 2, the next study [Wittek et al., 2016b] looks into the nature of citation behaviour over time to identify *entanglement* (without nonlocality) as a criterion of QL by Ncpol2spda

developed for PERICLES[9]. This software is a correlation analyzer, able to identify stronger-than-usual, non-classical correlations, which in turn tell us something about the incompleteness of the standard information representations for machine learning[10]. For the definition of entanglement and nonlocality please see the Stanford Encyclopedia of Philosophy[11].

INTRODUCTION

Citation and coauthor networks offer an insight into the dynamics of scientific progress where semantic content again interacts with evolving user behaviour. We can also view them as representations of a causal structure, a logical process captured in a graph. From a causal perspective, we can ask questions such as whether authors form groups primarily due to their prior shared interest, or if their favourite topics are “contagious” and spread through co-authorship. Such networks have been widely studied by the artificial intelligence community, and recently a connection has been made to nonlocal correlations produced by entangled particles in quantum physics -- the impact of latent hidden variables can be analyzed by the same algebraic geometric methodology that relies on a sequence of semidefinite programming (SDP) relaxations. Following this trail, below we treat our sample coauthor network as a causal graph and, using SDP relaxations, rule out latent homophily as a manifestation of prior shared interest leading to the observed patternedness. By introducing algebraic geometry to citation studies, we add a new tool to existing methods for the analysis of content-related social influences.

For a background, clarifying a line of argumentation by references, citations as a legacy mapping and orientation tool have been in use by knowledge organization for a long time. Their respective importance has led to the birth of new fields of study like scientometrics and altmetrics [Borgman & Furner, 2005; Zahedi et al., 2014; Cronin & Sugimoto, 2014], permeating funding decisions and ranking efforts [Vanclay, 2012; Hicks, 2012]. At the same time, citations embody scholarly courtesy as well as a form of social behavior, maintaining or violating norms [Cronin & Overfelt, 1994; Kaplan, 1965; Mitro, 1974; Gilbert, 1977; Ziman, 2000; Sandstrom, 2001; Börner et al., 2006]. Due to this, as is often the case when individual and social patterns of action are contrasted, one can suspect that factors not revealed to the observer of a single individual may point at underlying group norms when communities of individuals are scrutinized. To understand our own behavior as a species, it is important to detect any such influence.

Departing from earlier work [Ver Steeg & Galstyan, 2011], we turned to citation studies to find supporting evidence for signs of quantum-likeness in co-author behaviour. Our working hypothesis was that in citation patterns, a more fundamental layer would correspond to research based on shared interest between the author and her/his predecessors called *latent homophily*, whereas a more ephemeral second layer would link to current trends in science. Due to this, e.g. for a funding agency to find citation patterns going back to latent homophily would amount to better founded decisions, with such a pattern playing the role of a knowledge nugget. Consequently, ruling out latent homophily would correspond to a sieve or a filter, one important step in an anticipated workflow to dig for such nuggets by stratification in citations.

---

[9] The latest release is available at https://github.com/peterwittek/ncpol2sdpa/releases

[10] A digital preservation system is always made up of components: objects in the archives, users, usage scenarios, etc. The question is whether these components add up to more than just their sum. Ncpol2spda is a software tool for the macroscopic scanning of datasets to detect non-classical correlations. Such correlations belong to a class called *entanglement* in quantum mechanics, and indicate quantum-like behavior in the data. Ncpol2sdpa uses the method of sparse semidefinite programming relaxations for polynomial optimization problems of noncommuting variables to isolate them to exclude hidden variable theorems in networked data and verify the strength of observed correlations.

[11] http://plato.stanford.edu/entries/qt-entangle/

Related Research and Conceptual Clarifications

The notion of the citation network was famously developed by de Solla Price [de Solla Price, 1965] and since then it has evolved in many different directions. Incidentally, [Garfield et al., 1964] had already proposed the use of "Network Charts" of papers for the study of scientific history, but see also [Garfield et al., 2003] and [Garfield, 2009] for a newfound interest in algorithmic historiography. Although fruitful for analysis at a less aggregated level, these maps provide the possibility to visualize the network structure of single citing/cited papers of up to, say, the lower hundreds of papers before becoming too complex to overview. To remedy this, aggregated forms of citation networks have been developed, most notably bibliographic coupling [Kessler, 1963], 'co-mentions' of literary authors [Rosengren, 1968], and the more established concept of 'co-citation' of papers [Small, 1973]. Eventually, over time these aggregated forms of measurement were extended to analyse network structures of authors [McCain, 1986; White & Griffith, 1981]. By today, possibilities include the coverage of source titles and, for bibliographic coupling to reveal the networks based on address data such as department, institution and country, are limited only to the kind of structured data available in the database used for sampling [van Eck & Waltman, 2010 van Eck & Waltman 2014]. Common for many of these efforts is that the network structure is used to map or represent bibliometric data for descriptive purposes in visualization, while attempts at analyzing the relationships dynamically in more causal ways have not been considered to the same extent. A notable exception is [Bar-Ilan, 2008] for an overview of a third mode of aggregated co-studies, namely co-authorship studies that incorporate complex systems research and Social Network Analysis.

To address a different subject area, graphical models capture the qualitative structure of the relationships among a set of random variables. The conditional independence implied by the graph allows a sparse description of the probability distribution [Pearl, 2009]. Therefore by combining co-authorship and citation data we propose to view co-author and citation graphs as examples of such graphical models.

However, not all random variables can always be observed in a graphical model: there can be hidden variables. Ruling these out is a major challenge. Take, for instance, obesity, which was claimed to be socially contagious [Christakis & Fowler, 2007]. Is it not possible that a latent variable was at play that caused both effects: becoming friends and obesity? For the above assumption of latent homophily, [Ver Steeg & Galstyan, 2011] asks whether there is a limit to the amount of correlation between friends, at the same time being separable from other sources different from friendship. Or, do some smokers become connected because they had always smoked, or because copying an example may bring social rewards? To cite a methodological parallel, in quantum physics, the study of nonlocal correlations also focuses on classes of entanglement that cannot be explained by local hidden variable models -- these are known as Bell scenarios, initially stated as a paradox by Einstein, Podolsky and Rosen in their so-called EPR paper [Einstein et al., 1935].

As is well known, the EPR paper proposed a thought experiment which presented then newborn quantum theory with a choice: either supraluminal speed for signaling is part of nature but not part of physics, or quantum mechanics is incomplete. Thirty years later, in a modified version of the same thought experiment [Bell, 1964], Bell's Theorem suggested that two hypothetical observers, now commonly referred to as Alice and Bob, perform independent measurements of spin on a pair of electrons, prepared at a source in a special state called a spin singlet state. Once Alice measures spin in one direction, Bob's measurement in that direction is determined with certainty, as being the opposite outcome to that of Alice, whereas immediately before Alice's measurement Bob's outcome was only statistically determined (i.e., was only a probability, not a certainty). This is an unusually strong correlation that classical models with an arbitrary predetermined strategy (that is, a local hidden variable) cannot replicate.

Recently, algebraic geometry offered a new path to rule out local hidden variable models following from Bell's Theorem [Ver Steeg & Galstyan, 2011; Ma et al., 2015; Ver Steeg, 2015]. By describing probabilistic models as multivariate polynomials, we can generate a sequence of semidefinite programming relaxations which give an increasingly tight bound on the global solution of the polynomial optimization problem [Lasserre, 2001]. Depending on the solution, one might be able to reject a latent variable model with a high degree of confidence. In our case, Alice and Bob decide about references to be picked in complete isolation, yet their decisions, in spite of being independent from each other's, may be still correlated. If we identify the source of the shared state preceding their decisions as they make their choices, we can observe correlations between author pairs, and conclude that their patterns of citing behaviour cannot be explained *only* by the fact that they have always liked each other. In other words, experimental findings may rule out latent homophily in certain scenarios.

Citation Networks and Latent Homophily

To translate the above to experiment design, we must briefly discuss how latent homophily manifests in citation networks and why we want to restrict our attention to static models. We shall be interested in citation patterns of individual authors who have co-authored papers previously. Social 'contagion' means that authors will cite similar papers later on if they previously co-authored a paper. On the other hand, latent homophily means that some external factor -- such as shared scientific interest -- can explain the observed correlations on its own.

Given an influence model in which a pair of authors makes subsequent decisions, if we allow the probability of transition to change in between time steps, then arbitrary correlations can emerge. Static latent homophily means that the impact of the hidden variable is constant over time, that is, the transition probabilities do not change from one time step to the other. We restrict our attention to such models, this being a necessary technical assumption for the algebraic geometric framework. In practice, this means that an author does not get more or less inclined over time to cite a particular paper.

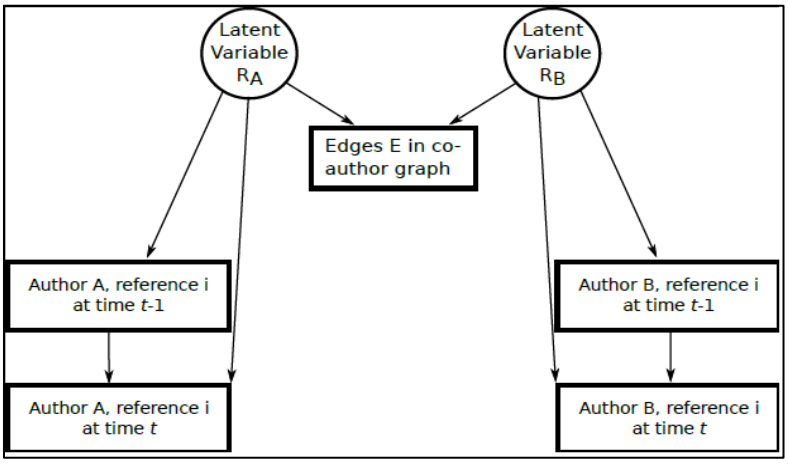

Fig. 4-1. Outline of the influence model. The latent variables R_A and R_B cause the edges in the co-author network and are also the sole influence in changes whether an author-reference pair changes in subsequent time steps.

A straightforward way to analyze correlations is to look at citation patterns between authors. Departing from a set of authors in an initial period, we can study whether the references an author makes influence the subsequent references of her or his coauthors as defined in the initial period. In this sense, we define a graph where each node is an author-reference. Two nodes are connected if the authors have co-authored a paper at some initial time step. A node is assigned a binary state ± 1,

reflecting whether that author-reference pair is actually present. The influence model is outlined in Fig. 4-1. Please find the mathematical details in the original publication [Wittek et al., 2016b].

DATASET

Data was collected from Web of Science, using the journal indices WoS-Extended, SSCI, and AHCI between 1945 and 2013 (Table 4-2). The collection consists of the full set of published items in 20 high impact journals found in the database. 43168 items were collected in total, comprising of 22784 articles (52.4%), 10270 book reviews (23.8%), 2325 editorial material papers (5.4%), and 1898 proceedings papers (articles) (4.4%).

The selection process was conducted by using four different journal rankings. The reason for using multiple source rankings was to minimize the impact of perspective, where, for example, the JCR ranking for Library and Information Studies (LIS) contains journals from the Information Systems subject area, however that would not count as (core) LIS journals by practitioners in the field. The ranking schemes used were JCR 2012, JCR 1997 (the oldest one found readily in the WoS platform), Google top publications (H5-Index), and Elsevier SCImago Rank 2012. Journal rank data and citation data were collected on January 20, 2014.

The inclusion of publication years 2013 and 2014 is not complete, since it is generally acknowledged that WoS has not received the underlying data until late spring the year after publication. Since the dataset is used for information based research and not for performance based evaluation, inclusion of as much as possible material was deemed more important than completeness.

To rank the journals, in all four lists the 20 top journals were scored from 20 to 1, so that the top journal earned 20 points and the last one earned 1 point. Then the points from each of the occurring journals in the four rankings were added and the journals were listed again based on their combined score for Table 4-2.

Table 4-2. The number of published entries, along with total number of citations, mean number of citations, and first year of inclusion in the WoS index.

Ord	Journal	Recs	Citations	Mean citations	Mean citations per year	First year
1	Journal of the American Society for Information Science and Technology	2494	22958	9.21	1.11	2001
	Journal of the American Society for Information Science	2977	39593	13.3	0.57	1970
	American Documentation	780	4347	5.57	0.11	1956
	Journal of Documentary Reproduction (United States)					
2	Journal of Informetrics	420	3714	8.84	1.69	2007
3	Scientometrics	3637	38202	10.5	0.94	1978
	Journal of Research Communication Studies	119	137	1.15	0.03	1978
4	Information Systems Research	649	25817	39.78	3.19	1994
5	MIS Quarterly	1071	70899	66.2	4.54	1981
6	College & Research Libraries	5156	12144	2.36	0.12	1956
7	Journal of the American Medical Informatics Association	4260	40687	9.55	0.95	1994
8	Library & Information Science Research	1209	6198	5.13	0.4	1984
	Library Research (United States)					
9	Annual Review of Information Science and Technology	550	7269	13.22	0.82	1966
10	Journal of Documentation	3700	18437	4.98	0.26	1945
11	Journal of Health Communication	1233	10570	8.57	0.99	1997
12	Journal of Information Science	1379	7802	5.66	0.29	1979
	Information Scientist (United Kingdom)					
	Institute of Information Scientists. Bulletin (United Kingdom)					
13	International Journal of Geographical Information Science	1299	14635	11.27	1.09	1997
	International Journal of Geographical Information Systems	311	6547	21.05	0.99	1991
14	Journal of Information Technology	612	5613	9.17	0.8	1993
15	Library Quarterly	4603	6200	1.35	0.07	1956
16	Journal of the Medical Library Association	1104	4275	3.87	0.44	2002
	Bulletin of the Medical Library Association	3639	10255	2.82	0.11	1956
17	Empty					
18	Arxiv Digital Libraries (cs.DL)					
19	Information & Management	1702	31902	18.74	1.52	1983
	Systems Objectives Solutions	63	274	4.35	0.13	
	Information Management	200	25	0.13	0	1983
	Management Datamatics (Netherlands)					
	Management Informatics (Netherlands)					
	IAG Journal (Netherlands)					
20	Reference Librarian					
	Total number of records	43167		11.53	0.88	

For every selected journal title, the title was run against the Ulrich's Periodicals Directory to identify title changes during the span of the journal's publishing history. In all, 33 versions of the titles were searched for in WoS. Of these, 24 titles were found in the database. The number of published entries, along with total number of citations, mean number of citations, and rst year of inclusion in the WoS index are presented in Table 4-2. The coauthor network has 45,904 nodes and 78,418 edges.

Experiment Design

There is an important distinction between reference and citation. While reference is a feature of the citing article in order to support an argument either as a fact or as a rhetorical tool, citation is a sign that indicates that a particular paper has been used, and therefore important metadata for DP. This binary nature of the citation, without any indication about how or why a document has been cited, turns it into a descriptor of the paper in some sense. Therefore we can conceive a bibliographic record with references as a combination of "internal metadata" describing the document itself, and "external metadata" by references for linkage, implementing citations.

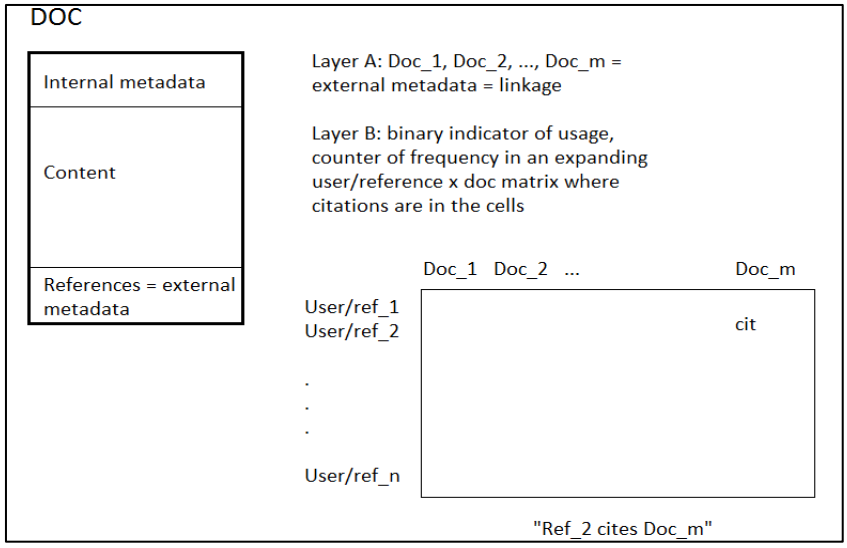

Fig. 4-2. A user/reference x document matrix, with citations/frequency counts in the cells. With those frequency counts, *tfidf* and all the weighting schemes from the VSM can be tested, plus the citation matrix can have its own semantics because any citations are context-dependent, where the context is the topic.

We decided to conduct an experiment with a semi-synthetic example to verify whether such a network of citations allows for the exclusion of latent hidden variables. For this case, to design a model of influence, the graph had to be directed, whereas a coauthor network is typically undirected. To establish directions in the graph, we considered a pairwise asymmetric relationship between authors in which one of the authors is `dominant'. To this end, we considered two alternatives:

1. The more dominant author is the one with more citations. As in our corpus every author pair has the same number of citations, this option was not viable and was therefore discarded;
2. The more dominant author has a higher degree in the graph of the coauthor network because he or she had more coauthors in the past. This enabled us to direct the graph.

We assumed that the network structure does not evolve over time. Taking the directed coauthor network graph in consideration, we assigned a state to each node, and set its value randomly with ± 1 with equal probability.

Once this initialization was done, we had to simulate influence. We randomly picked a pair, and the nondominant author copied the state of the dominant one. In a time step, we did M such random picks, where M is the number of edges. This gave sufficient opportunity for the graph to flip most of its nodes if necessary. We created two more time slices on top of the initial one. Using these time slices, we could calculate the statistics $P(A_{1:T}B_{1:T} \mid E = 1)$ with $T = 3$, where $E = 1$ meant that there was a directed edge from author A to author B.

With this random initialization, one can detect if, given a particular graph structure, there is a possibility of latent homophily at all. We used metaknowledge[12] to work with the citation network [McIlroy-Young & McLevey, 2015], Ncpol2sdpa[13] to generate the SDP relaxations [Wittek, 2015], and Mosek[14] to solve the SDP. The computational details are available online[15]. Taking the observables $O_j(A,B)$ as the indicator function and a level-3 relaxation of the Lasserre hierarchy, the SDP solver detects any dual infeasibility. In turn, such an infeasibility means that the sum-of-squares (SOS) decomposition does not exist and we can rule out latent homophily as the source of correlations with a high degree of confidence.

Static Latent Homophily in the Coauthor Network: Results and Discussion

We split the sample corpus into three periods with the following distribution (Table 4-3):

Table 4-3. Three periods in the sample longitudinal corpus with citation distributions.

Period	Number of papers
1945-1968	4104
1968-1991	12293
1991-2014	26770

As a joint probability distribution, one obtains 64 possible combinations of outcomes, because for each author and time period, the outcome is binary, and given two authors and three time periods, we obtain this number. We observe all possible outcomes on this sample. Using the same $O_j(A,B)$ observable as in the semi-synthetic example, i.e. the indicator function, and a level-3 relaxation of the Lasserre hierarchy, the SDP solver detects dual infeasibility, therefore we can rule out latent homophily as the single source of correlations.

Our result indirectly confirms that 'contagion' in the practice of citation is a distinct possibility. If citation patterns continue spreading, over time everybody will cite more or less the same papers. This in turn explains the phenomenon of Sleeping Beauties [Ke et al., 2015]: since dominant authors do not cite such articles, everybody else ignores them.

Secondly, we recall that in its simplest form, Bell's theorem states that no physical theory of local hidden variables can ever reproduce all of the predictions of quantum mechanics, i.e. it rules out such variables as a viable explanation of quantum mechanics. Therefore we hypothesized that if we can find entanglement in our data, with local hidden variables as their source ruled out, patterns in the sample must be quantum-like for non-obvious reasons. Ruling out Bell inequalities as the source of entanglement in our results points to such non-classical correlations at work in the dataset. On the other hand, nonlocality does not apply, reinforcing the QL verdict.

[12] http://networkslab.org/metaknowledge/
[13] https://pypi.python.org/pypi/ncpol2sdpa/
[14] https://mosek.com/
[15] http://nbviewer.jupyter.org/github/peterwittek/ipython-notebooks/blob/master/Citation_Network_SDP.ipynb

4.3. Description and Implementation of Dual Content Representations and Development of Quantum-based Models for Semantic Content Classification

4.3.1. Particle-like Index Terms, Drifts and "Gravity"

As a follow-up to work in progress in D4.4, in Experiment 3 we look at knowledge dynamics from a modelling perspective for another implementation of physics as a metaphor. This section is an extended version of a paper we submitted to Semantics 2016 [Darányi et al., 2016][16], and builds on the idea that the context-dependency of word and sentence meaning is expressed as their importance, while word and sentence similarity combined with importance constitutes conceptual fields expressed by lexemes. The interplay between word importance and similarity constitutes a source of semantics one can model by gravitation.

For a summary, in accessibility tests for digital preservation, over time one experiences drifts of localized and labelled content in statistical models of evolving semantics represented as a vector field. This articulates the need to detect, measure, interpret and model outcomes of knowledge dynamics. To this end we shall employ Somoclu, a high-performance machine learning algorithm for the training of extremely large emergent self-organizing maps for exploratory data analysis. The working hypothesis we present here is that the dynamics of semantic drifts can be modeled on a relaxed version of Newtonian mechanics called social mechanics. By using term distances as a measure of semantic relatedness vs. their PageRank values indicating social importance and applied as variable "term mass", gravitation as a metaphor to express changes in the semantic content of a vector field lends a new perspective for experimentation. From "term gravitation" over time, one can compute its generating potential whose fluctuations manifest modifications in pairwise term similarity vs. social importance, thereby updating Osgood's semantic differential. The dataset examined is the public catalog metadata of Tate Galleries, London.

INTRODUCTION

The evolving nature of digital collections comes with an extra difficulty: due to various but constant influences inherent in updates, the interpretability of the data keeps on changing. This manifests itself as **concept drift** [Wang et al., 2011] or **semantic drift** [Gulla et al., 2010; Wittek et al., 2015; Webb et al., 2016], the gradual change of a concept's semantic value as it is perceived by a community. Despite terminology differences, the problem is real and with the increasing scale of digital collections, its importance is expected to grow [Schlieder, 2010]. If we add drifts in cultural values as well, the fallout from their combination brings memory institutions in a vulnerable position as regards long term digital preservation. We illustrate this on a museum example, the subject index of the Tate Galleries, London.

In our example, semantic drifts lead to limited access by Information Retrieval (IR). The methodology we apply to demonstrate our point is vector field semantics by emergent self-organizing maps (ESOM) [Ultsch, 2005], because the interpretation of semantic drift needs a theory of update semantics [Veltman, 1996], integrated with a vector field rather than a vector space representation of content [Wittek et al., 2014; Wittek et al., 2015]. Further, given such content dynamics, we argue that for its modeling, one can fall back on tested concepts from classical (Newtonian) mechanics and differential geometry. For such a framework, e.g. similarity between objects or features can be

[16] See also: http://arxiv.org/abs/1608.01298

considered an attractive force, and changes over time manifest in content drifts have a quasi-physical explanation. The main contributions of this paper are the following:

1. A methodology for the detection, measurement and interpretation of semantic drift;
2. On drift examples, an improved understanding of how semantic content as a vector field "behaves" over time by falling back on physics as a metaphor;
3. As a consequence of the above, the concept of semantic potential as a combined measure of semantic relatedness and semantic importance.

We note in passing that the term frequency/inverse document frequency (*tfidf*) weighting scheme, typically used in mainstream vector space-based IR and ML, already implies semantic importance by the occurrence rate of index terms, indicating topical actuality. Our semantic potential reframes this component by combining it with similarity in a new way.

Background

Terminology

Evolving semantics (also often referred to as "**semantic change**" [Tury & Bieliková, 2006]) is an active and growing area of research into language change [Baker, 2008] that observes and measures the phenomenon of changes in the meaning of concepts within knowledge representation models, along with their potential replacement by other meanings over time. Therefore it can have drastic consequences on the use of knowledge representation models in applications. Semantic change relates to various lines of research such as ontology change, evolution, management and versioning [Meroño-Peñuela et al., 2013], but it also entails ambiguous terms of slightly different meanings, interchanging shifts with drifts and versioning, and applied to concepts, semantics and topics, always related to the thematic composition of collections [Yildiz, 2006; Uschold, 2006; Klein & Fensel, 2001]. A related term is semantic decay as a metric: it has been empirically shown that the more a concept is reused, the less semantically rich it becomes [Pareti et al., 2015]. Though largely counter-intuitive, this derivation is based on the fact that frequent usage of terms in *diverse* domains leads to relaxing the initially strict semantics related to them. The opposite would hold if a term was persistently used within a single domain (or to a great extent similar domains), which would lead to its gradual specialization and enrichment of its semantics.

Related Research

Here we mention four relevant directions, all of them contributors to our understanding of a complex issue in their overlap.

Temporality and Advanced Access

By advanced access to digital collections we mean the spectrum of automatic indexing, automatic classification, Information Retrieval (IR), and information visualization. All of the aforementioned can have a temporal aspect, increasingly being addressed by current research. Examples for temporal IR include [Alonso et al., 2011; Chang et al., 2013], for web dynamics and visualization, see e.g. [Dubinko et al., 2006; Adar et al., 2008; Sharapenko et al., 2005]. A related but separate research area for the above is in the overlap of cultural heritage and IR [Koolen et al., 2009; de Jong et al., 2005].

Vector Space vs. Vector Field Semantics

For an IR model to be successful, its relationship with at least one major theory of word meaning has to be demonstrated. With no such connection, meaning in numbers becomes the puzzle of the ghost in the machine. For the vector space IR model (VSM) – underlying many of today's competitive IR products and services – such a connection can be demonstrated; for others like PageRank [Page &

Brin, 1998], the link between graph theory and linear algebra leads to the same interpretation. Namely, in both cases, the theory of word semantics cross-pollinating numbers with meaning is of a contextual kind, formalized by the distributional hypothesis [Harris, 1968], which posits that words occurring in similar contexts tend to have similar meanings. As a result, the respective models can imitate the field-like continuity of conceptual content. However, unless we consider the VSM roots of both the probabilistic relevance model[17] and its spinoffs including BM25[18], such a link is still waiting to be shown between probability and semantics [Frommholz et al., 2010].

Although several attempts exist to this end [Turney & Pantel, 2010; Pulman, 2013], a brief overview should be helpful. Looking for a good fit with some reasonably formalized theory of semantics, two immediate questions emerge. First, can the observed features be regarded as entries in a vocabulary? If so, distributional semantics applies and, given more complex representations, other types may do so as well [Wittek et al., 2013b]. The second question is, do they form sentences? For example, one could regard a workflow (process) a sentence, in which case compositional semantics applies [Coecke et al., 2010; Sadrzadeh & Grefenstette, 2011]. If not, theories of word semantics should be considered only. Below we shall depart from this assumption.

Notwithstanding the fact that vector space in its most basic form is not semantic, its ability to yield results which make sense goes back to the fact that the context of sentence content is partially preserved even after having eliminated stop words which are useless for document indexing. This means that Wittgenstein's contextual theory of meaning ("Meaning is use") holds [Wittgenstein, 1963], also pronounced by the distributional hypothesis. This is exploited by more advanced vector based indexing and retrieval models such as Latent Semantic Analysis (LSA) [Deerwester et al., 1990] or random indexing [Kanerva et al., 2010], as well as by neural language models, ranging from the Simple Recurrent Networks, and their very popular flavour, Long Short-Term Memory (LSTM) [Hochreiter & Schmidhuber, 1997] and the recently proposed Global Vector for Word Representation (GloVe) [Pennington et al., 2014], which are currently considered to be the state-of-the-art approach to text representation. However, we should also remember another approach paraphrased as "Meaning is change", namely the stimulus-response theory of meaning proposed e.g. by Bloomfield[19] in anthropological linguistics and Morris[20] in behavioral semiotics, plus the biological theory of meaning [Uexküll & Kriszat, 1956]. These authors stress that the meaning of an action is in its consequences. Consequently word semantics should be represented not as a vector space with position vectors only, but as a dynamic vector field with both position and direction vectors [Wittek et al., 2014].

Linguistic "Forces"

As White suggests, linguistics, like physics, has four binding forces [White, 2002]:

1. The strong nuclear force, which is the strongest "glue" in physics, corresponds to word uninterruptability (binding morphemes into words);
2. Electromagnetism, which is less strong, corresponds to grammar and binds words into sentences;
3. The weak nuclear force, being even less strong, compares to texture or cohesion (also called coherence), binding sentences into texts;
4. Finally gravity as the weakest force acts like intercohesion or intercoherence which binds texts into literatures (i.e. documents into collections or databases).

Mainstream linguistics traditionally deals with Forces 1 and 2, while discourse analysis and text linguistics are particularly concerned with Force 3. The field most identified with the study of Force 4

[17] Because it departs from a "binary index descriptions of documents", see [Robertson & Spärck Jones, 1976].
[18] See p. 339 in [Robertson & Zaragoza, 2009].
[19] https://en.wikipedia.org/wiki/Leonard_Bloomfield
[20] https://en.wikipedia.org/wiki/Charles_W._Morris

is information science. As the concept of force implies, referring here to attraction, it takes energy to keep things together, therefore the energy doing so is stored in agglomerations of observables of different kinds in different magnitudes, and can be released from such structures. A notable difference between physical and linguistic systems is that extracting work content, i.e. “energy" from symbols by reading or copying them does not annihilate symbolic content.

Looking now at the same problem from another angle, in the above and related efforts, energy inherent in all four types can be the model of e.g. a Type 2, i.e. electromagnetism-like attractive-repulsive binding force such as lexical attraction, also known as syntactic word affinity [Beeferman et al., 1997] or sentence cohesion, such as by modeling dependency grammar by mutual information [Yuret, 1998]. In a text categorization (TC) and/or IR setting, a similar phenomenon is term dependence based on their co-occurrence.

Semantic Kernels and “Gravity”

A radial basis function (RBF) kernel, being an exponentially decaying feature transformation, has the capacity to generate a potential surface and hence create the impression of gravity, providing one with distance-based decay of interaction strength, plus a scalar scaling factor for the interaction, i.e. $K(x,x') = \exp(-\gamma||x - x'||^2)$ [Moschitti, 2010]. We know that semantic kernels and the metric tensor are related [Williams et al., 2005], hence some kind of a functional equivalent of gravitation shapes the curvature of classification space [Amari & Wu, 1999; Eklund, 2016]. At the same time, gravitation as a classification paradigm [Peng et al., 2009] or a clustering principle [Aghajanyan, 2015] is considered as a model for certain symptoms of content behavior.

Working Hypothesis and Methodology

In order to combine semantics from computational linguistics with evolution, we select the theory of semantic fields [Trier, 1934] and blend it with multivariate statistics plus the concept of fields in classical mechanics to bring it closer to Veltman’s update semantics [Veltman, 1966], and to enable machine learning. Our working hypothesis for experiment design is as follows:

- Semantic drifts can be modeled on an evolving vector field as suggested by [Wittek et al., 2015; Wittek et al., 2014];
- To follow up on the analogy from semantic kernels defining the curvature of classification space and let this curvature evolve, Newton’s universal law of gravitation can be adapted to the idea of the dynamic library [Salton, 1975]. To this end, we model similarity by $F = Gm_1m_2/r^2$, with term dislocations over timesteps stored in distance matrices. Ignoring G, we shall use the PageRank value of index terms on their respective hierarchical levels for mass values. Since force is the negative gradient of potential, i.e. $F(x) = -dU/dx$, we can compute this potential surface over the respective term sets to conceptualize the driving mechanism of semantic drifts;
- The potential following from the gravity model manifests two kinds of interaction between entries in the indexing vocabulary of a collection. Over time, changes in collection composition lead to different proportions of semantic similarity vs. authenticity between term pairs, expressed as a cohesive force between features and/or objects.

For the analysis of context-dependent index term correlations we use a high-performance machine learning algorithm called Somoclu (“Self-Organizing Maps Over a Cluster”). This tool is primarily meant for training extremely large emergent self-organizing maps on supercomputers, but is also the fastest implementation running on a single node for exploratory data analysis [Wittek et al., 2015a]. The mathematical details are provided in [Wittek et al., 2015b][21].

[21] The latest release is available at https://github.com/peterwittek/somoclu/releases/tag/1.6.2

Drift Detection

The task of drift detection, measurement and interpretation is carried out in three basic steps as follows:

Step 1: Somoclu maps the high-dimensional topology of multivariate data to a low-dimensional (2-d) embedding by ESOM. The algorithm is initialized by LSA, Principal Component Analysis (PCA), or random indexing, and creates a vector field over a rectangular grid of nodes of an artificial neural network (ANN), adding continuity by interpolation among grid nodes. Due to this interpolation, content is mapped onto those nodes of the ANN that represent best matching units (BMUs).

Step 2: Clustering over this low-dimensional topology marks up the cluster boundaries to which BMUs belong. Their clusters are located within ridges or watersheds watersheds [Ultsch, 2005; Tosi et al., 2014; Lötsch & Ultsch, 2014]. Content splitting tendencies are indicated by the ridge wall width and height around such basins so that the method yields an overlay of two aligned contour maps in change, i.e. content structure vs. tension structure. In Somoclu, nine clustering methods are available. Because self-organizing maps (SOM), including ESOM, reproduce the local but not the global topology of data, the clusters should be locally meaningful and consistent on a neighborhood level only.

Step 3: Evolving cluster interpretation by semantic consistency check can be measured relative to an anchor (non-shifting) term used as the origin of the 2-d coordinate system, or by distance changes from a cluster centroid, etc. In parallel, to support semiautomatic evaluation, variable cluster content can be expressed for comparison by histograms, pie diagrams, etc.

Dataset and Experiment Design

Tate Subject Index

Tate holds the national collection of British art from 1,500 to the present day and international modern and contemporary art. The collection embraces all media, from painting, drawing, sculpture and prints to photography, video and film, installation and performance. The 19th century holdings are dominated by the Turner Bequest with cca 30,000 works of art on paper, including watercolors, drawings and 300 oil paintings. The catalog metadata for the 69,202 artworks that Tate owns or jointly owns with the National Galleries of Scotland are available in JSON format as open data[22]. Out of the above, 53,698 records are timestamped. The artefacts are indexed by Tate's own hierarchical subject index which has three levels, from general to specific index terms.

Analysis Framework Description

To study the robust core of a dynamically changing indexing vocabulary, we filtered the dataset for a start. As statistics for the Tate holdings show two acquisition peaks in 1796-1844 (33,625 artworks) and 1960-2009 (12,756 artworks), we focused on these two periods broken down into 10 five-years epochs each, with altogether 46,381 artworks. In the 19th century period, subject index level 1 had 22 unique general index terms (21 of them persistent over ten epochs), level 2 had 203 unique intermediate index terms (142 of them persistent), and level 3 had 6,624 unique specific index terms (225 of them persistent). In the 20th century period, level 1 had 24 unique terms (22 of them persistent), level 2 used 211 unique terms (177 of them persistent), and level 3 had 7,536 unique terms (288 of them persistent over ten epochs). Fig. 4-3 explains the analysis framework. Table 4-4 displays a sample entry from the subject index[23].

Following text pre-processing, which included the application of tokenization and stop-word removal on all three levels of concepts in the subject index, adjacency matrices and subsequently graphs were

[22] https://github.com/tategallery/collection
[23] http://www.tate.org.uk/art/artworks/turner-self-portrait-n00458

created using the co-occurrence of the terms in the artworks as undirected, weighted edges. These matrices were then used to extract an importance measure for each term by employing the PageRank algorithm, and to create ESOM maps using the Somoclu implementation.

For each of the 80 epochs (2 periods x 4 levels x 10 epochs), the ESOM's codebook was first initialized by employing PCA with randomized SVD, which was then used for mapping the high-dimensional co-occurrence data to an ESOM with a toroid topology. The results were represented on the two-dimensional projection of the toroid using different granularities according to the indexing level (20x12 = level 1, 40x24 = level 2, 50x30 = level 3, 60x40 = all levels together). Introducing the least displaced term per indexing level over a period as an anchor against which all term drifts on that level could be measured, we tracked the tension vs. content structure of evolving term semantics and evaluated the resulting term clusters for their semantic consistency.

The input matrices were processed by Somoclu as described above and the codebook of each ESOM was clustered using the affinity propagation algorithm [Frey & Dueck, 2007]. The results were tested for robustness by hierarchical cluster analysis (HCA), using Euclidean distance as similarity measure and farthest neighbor (complete) linkage to maximize distance between clusters, keeping them thereby both distinct and coherent. The ESOM-based cluster maps expressed the evolving semantics of the collection as a series of 2-dimensional landscapes over 10 epochs times two periods.

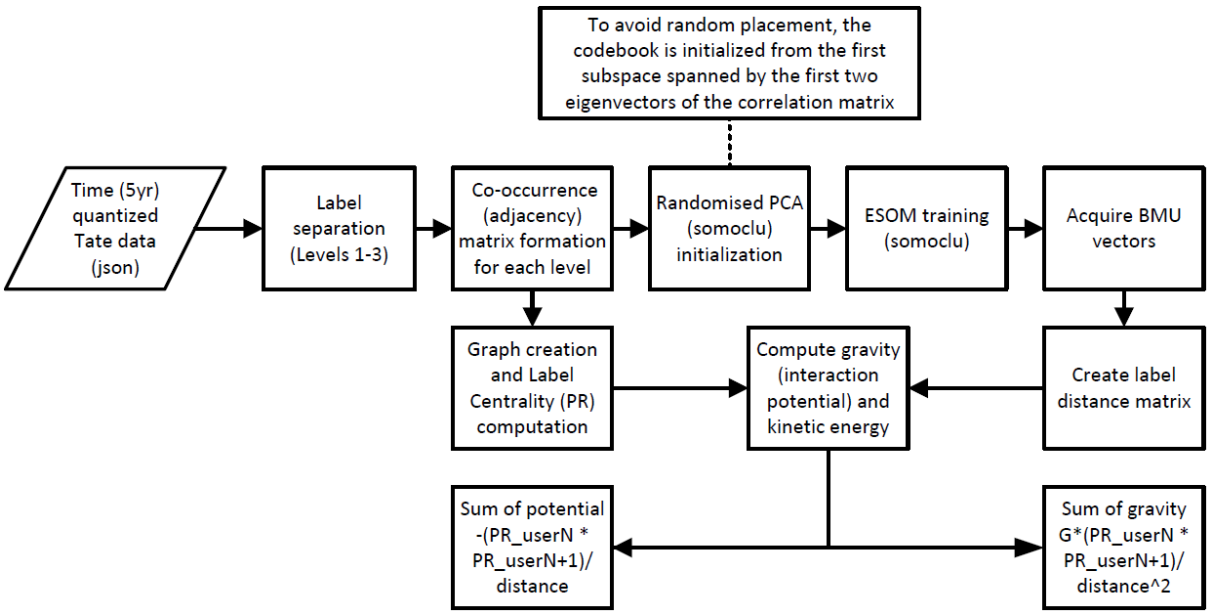

Fig. 4-3. Steps of the Tate processing workflow.

Table 4-4. Sample index terms describing a Turner self-portrait.

level 1 (general)	***level 2 (intermediate)***	***level 3 (specific)***
Objects	Clothing and personal effects	Cravat
People	Adults	Man
Named individuals	Turner, Joseph Mallord William	-
Portraits	Self-portraits	-
Work and occupations	Arts and entertainment	Artist, painter

Term drift detection, measurement and interpretation were based on these maps. To enable drift measurement, we generated a parallel set of maps with the term of greatest importance over all periods as its anchor point. Importance was defined by the Reciprocal Rank Fusion coefficient [Cormack et al., 2009] which combined the PageRank values of each term over all periods. This relative location was used for the computation of respective term-term distance matrices over every epoch of each period. Term dislocations over epochs were logged, recording both the splits of term clusters mapped onto a single grid node in a previous epoch, or the merger of two formally independent nodes labelled with different terms into a single one. These splits and merges were used to define the drift rate and subsequently the stability of the lexical field.

Finally, as per the second point of the working hypothesis, the gravity and potential surfaces for every epoch were computed. When computing gravity and potential, the property of mass was expressed via each term's PageRank score and the distance by measuring the normalized (sum to 1) Euclidean distance between the corresponding BMU vectors.

Results

Index term drift detection, measurement and evaluation were based on the analysis of ESOM maps, leading to drift logs on all indexing levels. Parallel to that, covering every time step of collection development, we also extracted normalized histograms to describe the evolving topical composition of the collection, and respective pie charts to describe the thematic composition of the clusters. Further, to check cluster robustness, hierarchical cluster analysis (HCA) dendrograms were computed for term-term matrices, also compared with those from term-document matrices. On one hand, these gave us a detailed overview of semantic drift in the analyzed periods. On the other hand, the observed dynamics could be modeled on the gravitational force and its generating potential.

A more detailed report would go beyond the opportunities of this report. However, some key indications were the following.

Semantic Drifts

Content mapping means that term membership for every cluster in every time step is recorded and term positions and dislocations over time with regard to an anchor position are computed, thereby recording the evolving distance structure of indexing terminology. This amounts to drift detection and its exact measurement. Adding a drift log results in extracted lists of index terms on all indexing hierarchy levels plus their percentage contrasted with the totals. Drifts can be partitioned into splits and merges. In case of a split, two concept labels that used to be mapped on the same grid node in one epoch become separated and tag two nodes in the next phase, while for a merge, the opposite holds. From an IR perspective splits decrease recall and merges decrease precision, limiting the quality of access; from a LTDP perspective they indicate at-risk indexing terminology.

Splits and merges were listed by Somoclu for every epoch over both periods. For instance a sample semantic drift log file recorded that due to new entries in the catalog in 1796-1800, by 1800 on subject index level 2, the term art was separated from works, as much as scientific was from measuring, whereas monuments, places and workspaces were merged, i.e. mapped onto the same coordinates. Therefore, based on the same subject index terms, anyone using this tool in 1800 would have been unable to retrieve the same objects as in 1796.

In a vector field, all the terms and their respective semantic tags are in constant flux due to external social pressures, such as e.g. new topics over items in the collection due to the composition of donations, fashion, etc. Without data about these pressures quasi embedding and shaping the Tate collection, the correlations between social factors and semantic composition of the collection could not be explicitly computed and named. Still, some trends could be visually recognized over both series of maps, going back to their relatively constant semantic structure where temporary content

dislocations did not seriously disturb the relationships between terms, i.e. neighboring labels tended to stick with one another, such as “towns, cities, villages” vs. inland and natural. In other words, the lexical fields as locally represented by Somoclu remained relatively stable.

The stability of these fields was measured in terms of drift rates which were computed by detecting the splits and merges that happened to the BMUs (e.g. Fig. 4-4). Specifically, we were not looking at the distance they travelled, rather at the fact that they formed or joined or moved away from a cluster (BMU) in between epochs.

Overall, in this particular collection, splits between level 1 concepts took place occasionally, whereas both splits and merges occurred on indexing levels 2-3 on a regular basis. The drift rate was increasingly high: for level 2 index terms, it was 19-22% in the 1796-1845 period vs. 15-27.5 % in 1960-2009, whereas for level 3 terms it was 29-57% (1796-1845) vs. 54-61 % (1960-2009). These percentages suggest that the more specific the subject index becomes, the more volatile its terminology, especially with regard to modern art.

Content vs. Tension Structure and Content Dynamics

To describe the composition of the social tensions shaping this collection, one can compare e.g. the level 2 indexing vocabularies for both periods. In general, this is where one witnesses the workings of language change, part producing new concepts, part letting certain index terms decay. E.g. focus is shifting from a concept to its variant (e.g. nation to nationality), a renaissance of interest in the transcendent beyond traditional notions of religion and the supernatural (occultism, magic, tales), fascination for the new instead of the old, or a loss of interest in royalty and rank. Toys and concepts like tradition, the world, culture, education, films, games, electricity and appliances make a debut in art. A representation of such tendencies in content change with manifest tensions is visualized in Fig. 4-4. Here, tendency means a projected possible, but not necessarily continuous, trend - should the composition of the collection continue to evolve over the next epoch like it used to develop over the past one, the indicated splits and merges would be more probable to form new content agglomerations than random ones.

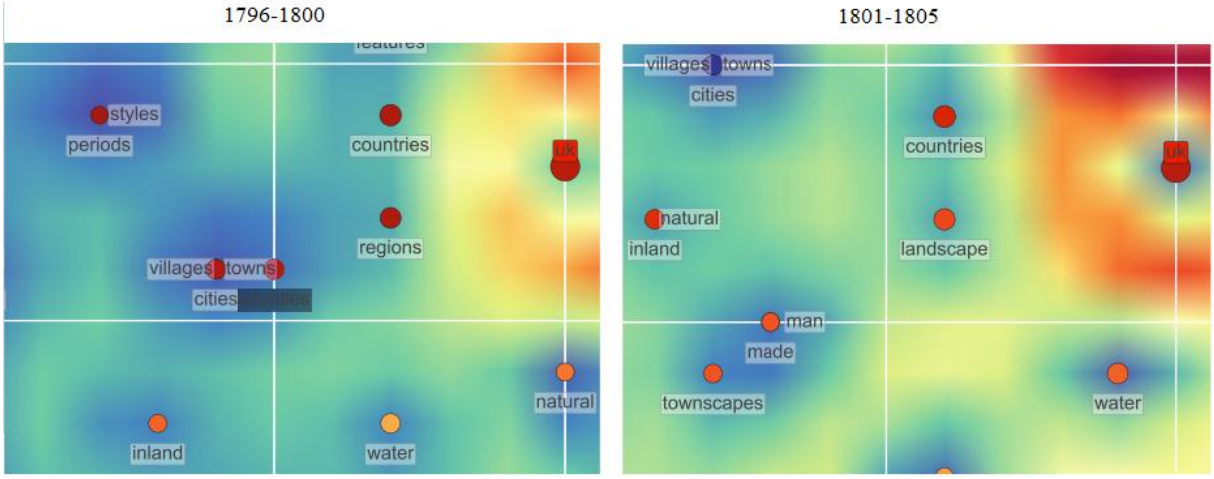

Fig. 4-4. Excerpt from the tension vs. content structure changes in the level 2 (intermediate) index term landscape in 1796-1805. Blue basins host content, brown ridges indicate tensions. Whereas *towns, cities, villages* remain merged over both epochs, *inland* and *natural* become merged within the same basin.

As we were left with the impression that in a statistically constructed vector field of term semantics drifts are the norm and not the exception, to account for such dynamics we computed a series of epoch-specific gravitational fields and their generating potential for a first overview. With BMU vector distances between term pairs and their PageRank values for “term mass”, both types of

surfaces expressed the interplay between semantic similarity and term importance in a social perspective (Fig. 4-5). This potential can be seen as the conceptual consequence of the semantic differential [Osgood et al., 1957], a forerunner to modern latent semantic methods. The semantic potential, in turn, suggests that physics as a metaphor is useful because it yields new, helpful concepts to model the dynamics of meaning, itself important for knowledge organization and knowledge management.

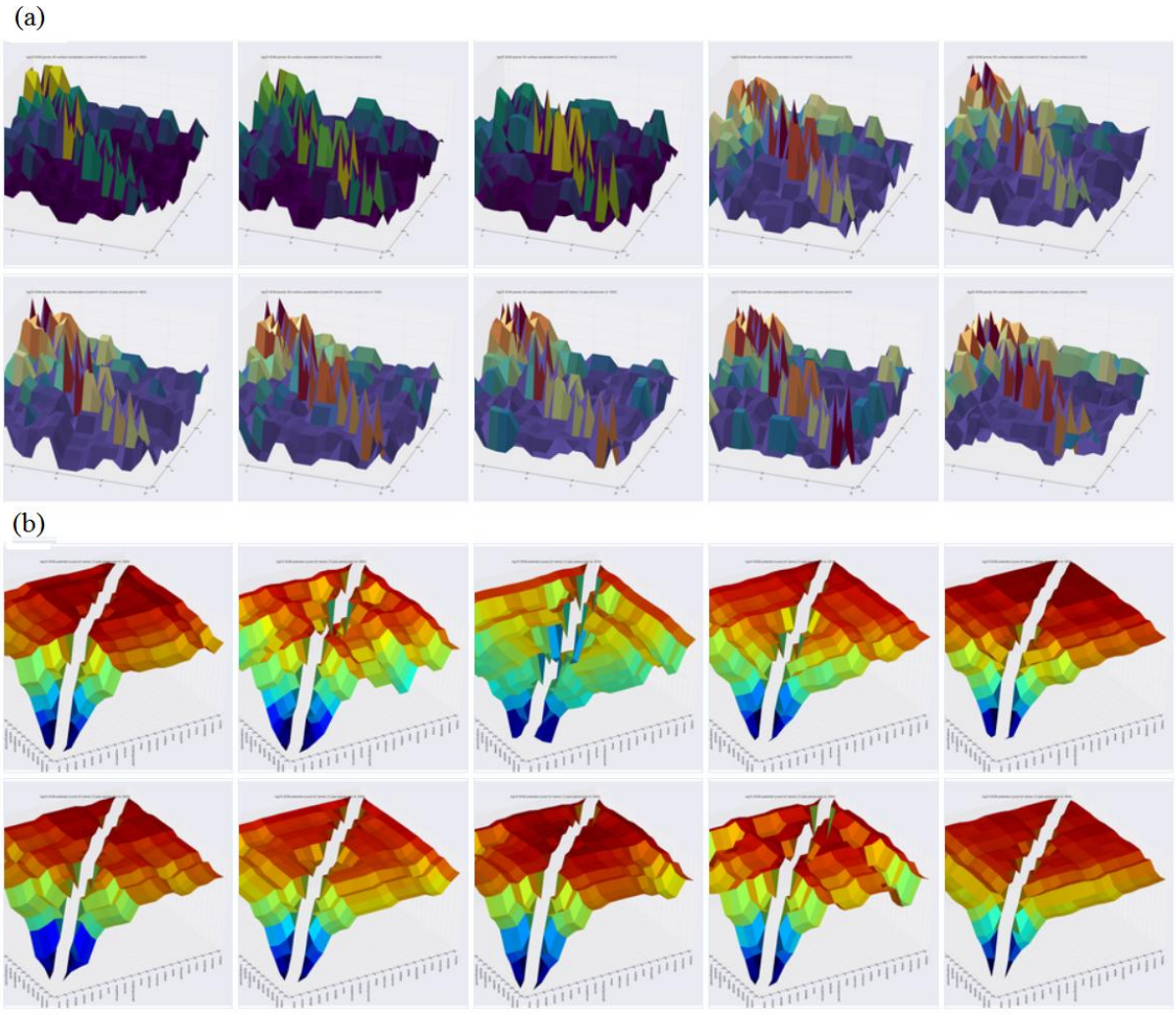

Fig. 4-5. (a) Changes in the top [level 1] conceptual layer of the Tate indexing vocabulary in 1796-1845, sampled every 5 years, modeled on a gravitational field. Gravitational force is the negative gradient of the corresponding potential. (b) Respective changes in the underlying potential field. Extreme values indicate semantically related term pairs with respectively high social status expressed by PageRank.

Moreover, based on the epoch-specific drift data, we were able to compute the kinetic energy (KE) vs. interaction potential (IP) of index terms. The procedure is as follows:

1. In order to prove the existence of an external potential and to calculate its change over time, we had to disprove the conservation of the sum between the interaction potential/gravity ($IP = cm_i\, m_j/r^2$) and that of the kinetic energy ($KE = \frac{1}{2} m_i v_i^2$) in the system. Since we know that total energy is conserved, KE+IP must be constant between time periods. If it is not, then there must be an external potential (EP), which affects the defined closed system and which we label as semantic “dark matter”;

2. Where the IP is concerned, we identify the mass m_i of a term i as its respective PageRank value for a particular timeslot, and the distance r between two masses as the Euclidean distance between their two Best Matching Units (BMUs). The BMUs are the codebook vectors from the pool that have the minimum distance to the input vectors (In our case, the adjacency vectors representing the co-occurrence between the terms in the dataset). It should be noted that the distance matrix constructed by evaluating the distance between all the terms' BMUs is incremented by a value of one in order to compensate for terms with identical BMUs. Such cases exist and signify an increased semantic similarity between the terms. In addition, we disregard the gravitational constant in the interaction potential/gravity formula, i.e. we set $G = c = 1$, which guarantees that we treat gravitation as a metaphor.

Regarding the KE, mass is identified as above and the velocity v_i of a term is defined as its rate of dislocation on the ESOM across time periods. In this case, dislocation is measured by calculating the Euclidean distance between the term's consecutive ESOM BMUs over time.

Having calculated the total energy for all the terms (sum of KE and IP) for each timeslot and having witnessed a serious fluctuation over time, according to the law of conservation of energy for closed/isolated systems, we concluded that there must be an additional potential that affects the system. For the all-inclusive summaries over both measurement periods, see Fig. 4-6.

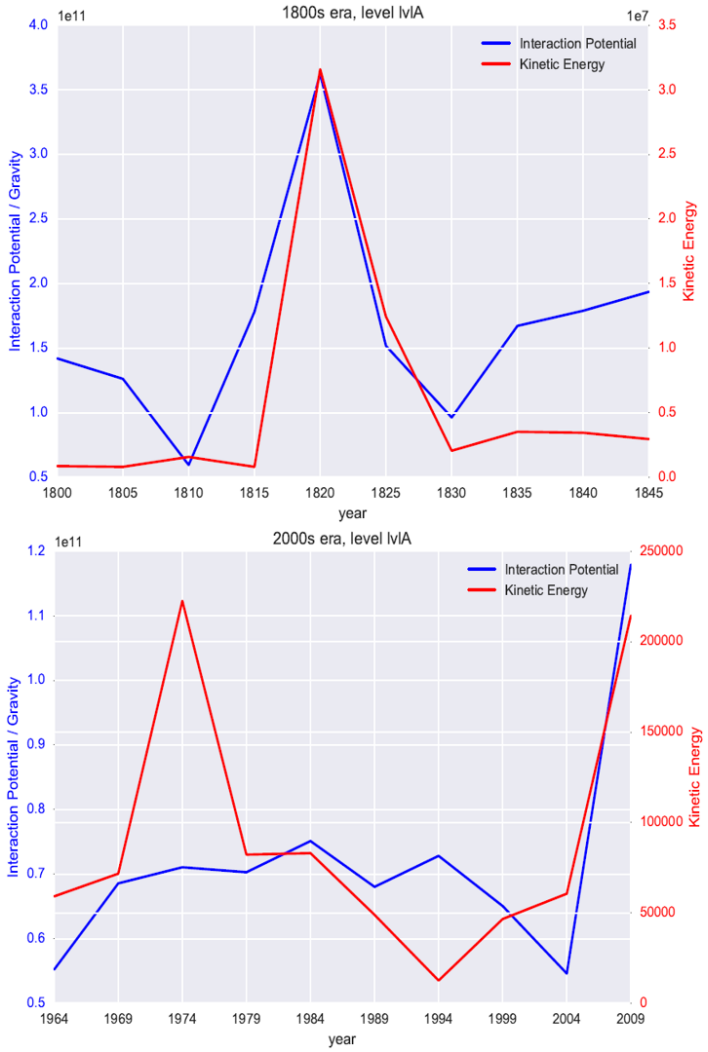

Fig. 4-6. External potential of the Tate collection in the two measurement periods.

The importance of this figure is that for a closed system of interactions, i.e. conservative forces, the sum of KE + IP = EP at any measurement point is constant, which is not the case here. That has two immediate implications:

1. The EP component represents a steady influence from the social environment embedding the system, accompanying the drifts and contributing to their existence by repeated system state updates. By adding new material to existing collections, these updates add new terms to the indexing vocabulary and/or modify the proportions of pre-existing ones, and thereby manifest social pressures with an impact on a system state;
2. The non-constant EP is indirect evidence for the distributional hypothesis used to construct semantic spaces and being responsible for their meaningfulness. As the concept implies, in spite of having been exploited for decades now, it was never confirmed to leave its hypothetic status.

Further, since KE + IP = EP is related to the total energy content of a system state called its Hamiltonian, we can continue our line of thought and focus on the computation of this component. The implications of this will be detailed in Section 4.3.4.

4.3.2. Index Terms Between Classical and Quantum Mechanics

The Role of the Hamiltonian in Evolving Document Collections: a Link to QL Systems

Taking into consideration that $H = T + V$ is the Hamiltonian equation we want to interpret – this is related to KE + IP = EP above –, where H is the Hamiltonian operator, T is the kinetic energy and V is the potential energy of a system, respectively, we argue that $AA^T = H$, that is, we treat the term co-occurrence matrix as the description of the total energy of the system. Thereby we also assume that our system is a conservative one. The same assumption was made by quantum clustering (QC) and dynamic quantum clustering (DQC) [Horn & Gottlieb, 2001; Weinstein & Horn, 2009; Di Buccio & Di Nunzio, 2011].

Any update of AA^T results in an $A'A'^T$ state with its corresponding V' potential energy, whereas the difference between any two consecutive V' goes back partly to changes in document collection content reflected by different index term occurrence rates (a.k.a. term frequency), partly to changes in the proportion of referential meaning added to H by sense definitions and sense relations of index terms. Both T and V can be analyzed by comparing consecutive spectral decomposition of the same index term over periods.

It is key to the understanding of V to remember that the semantic interpretation of both A and AA^T goes back to term occurrences in context, and thereby to the distributional hypothesis of word meaning [Harris, 1970]. However, taking a broader view of the issue, it is clear that at least one more factor, i.e. referential meaning must play a role in interpreting the above matrices as well. Namely the reason why terms in a particular context co-occur goes back to their ontological meaning, in a referential relation with their occurrences in sentences. This external, hidden contribution can be measured e.g. by the inverse relationship between the number of intensions (features) of a word vs. its extensions (cardinality of the set of its examples) [Carnap, 1947].

Considering our index terms as particles in state space, a many-body Hamiltonian of nonrelativistic, interacting particles can be written as:

$$H = \sum_{i=1}^{N} \frac{P_i^2}{2M_i} + \frac{1}{2\sum_{i,j=1}^{N} V_{ij} Q_i Q_j}$$

where V_{ij} is a symmetric positive semidefinite matrix. The sum over the particle interactions goes over all indices, which implies that the interaction length is infinite. This description assumes that

there is no external potential. If there is, the picture is more complicated, with the potential energy depending on both the external potential and the interaction potential[24].

$$H = \sum_{i=1}^{N} \frac{P_i^2}{2M_i} + V(Q_1, \ldots, Q_N, t)$$

In some cases, this reduces to

$$H = \sum_{i=1}^{N} \left(\frac{P_i^2}{2M_i} + V(Q_i, t) \right) + \frac{1}{2\sum_{i,j=1}^{N} V_{ij} Q_i Q_j}$$

that is, the impact of the external potential acts on individual particles, and we consider the interaction term separately.

Semantic Content and Particle-Wave Duality

Here we take a detour to explain the implications of the above. To repeat our working hypothesis, word meaning can be expressed as "energy" in index terms [Wittek & Darányi, 2011; Darányi & Wittek, 2012], because – as we have seen above in Section 4.3.1 – semantic content located in vector space generates a potential with energy minima on a potential surface. Such content constitutes regions with different semantic density [Mihalcea & Moldovan, 1998] so that both concepts and categories as their combinations are modelled by the above minima. Mathematical "energy" and ML are related, the latter often being based on minimizing a constrained multivariate function such as a loss function. Concepts in feature space "sit" at global energy minima, representing the cost of a classification decision as an energy minimizing process. This suggests that ML must identify concepts with such minima, and since potential energy in physics is carried by a field or a respective topological mapping, concepts naturally have something to do with energy as work capacity.

The solution we proposed in [Wittek et al., 2014] was to use emergent self-organizing maps to generate an artificial semantic field. The resulting space in this regard is a two dimensional surface, and the vector field associated with the points on the surface is a high-dimensional one. Combining this approach with earlier semantic models using the Hamiltonian of a quantum system [Darányi & Wittek, 2012], we want to see a dynamic model of language change to emerge.

To revisit energy-like objectives in ML, supervised learning algorithms measure the difference between target labels and the predictions of the model being trained. The goal is to minimize the difference: in such a scenario, we may regard the objective as "energy", and we look for a global minimum. This is not always the case, as error on the training sample does not necessarily imply a good generalization performance on unseen examples, as we know it from the theory of structural risk minimization. Hence, for instance, support vector machines do not fit this paradigm, but feedforward neural networks and certain types of boosting algorithms do.

Some unsupervised algorithms also seek a minimum on a high-dimensional surface, which, again, we may treat as a metaphor of energy. Examples include Hopfield networks, which map to an Ising Hamiltonian, or dynamic quantum clustering, where data instances are rolled along a potential surface to local minima. Dynamic quantum clustering is more direct in using energy as a metaphor [Weinstein & Horn, 2009]. It takes as the ground solution for the generic Hamiltonian of the Schrödinger equation:

$$H\psi = \big(T + V(x)\big)\psi = E_0\psi$$

where H is the Hamiltonian, T is the kinetic energy, V is the potential energy, and E_0 is the ground energy level. The algorithm evolves the Hamiltonian to identify the clustering structure by tracking

[24] https://en.wikipedia.org/wiki/Hamiltonian_(quantum_mechanics)

the expectation values of the position operator. According to Ehrenfest's theorem, the expectation values of the position operator obey their corresponding classical equations of motion, i.e.

$$\frac{d^2\langle x(t)\rangle}{dt^2} = \langle \psi(t) \vee \nabla V \vee \psi(t)\rangle$$

That is, the centre of each wave packet rolls towards the nearest minimum of the potential according to Newton's law of motion, i.e. a spatiotemporally limited bunch of waves behaves like a particle [Daránvi & Wittek, 2012].

To sum up developments so far, taking the spectrum of the Hamilton operator *H* in a finite dimensional space, we conjectured that index terms are associated with a set of eigenvalues, giving them a spectral signature [Wittek & Daránvi, 2011]. The eigenvalues corresponded to the different senses of a word, where a higher level energy state was more unlikely to be occupied. Following a different train of thought, Daránvi and Wittek (2012) studied the kinetic term of the Hamiltonian, *T*, to identify words with weights, and derive dynamics through Ehrenfest's theorem [Daránvi & Wittek, 2012]. What has been missing so far is the potential term in the Hamiltonian, which is also the most complex one. By Somoclu, we ventured a step towards defining a potential field by interpolating the distributional semantic description of term vectors, whereas above, we have also identified both the kinetic and the potential terms of the Tate Hamiltonian.

By a somewhat different track [Daránvi & Eklund, 2007] applied QC [Horn & Gottlieb, 2001], a probabilistic method, for the visualization of contour maps as a result of term and/or document classification. In the first step, just like with Somoclu, one starts with dimension reduction techniques to limit the list of word forms extracted from the records to only those which impact the index term distribution to the greatest extent. Then a contour map of the documents indexed by these terms is generated, so that the ”longitude” and the ”latitude” of the map are computed by SVD, whereas its ”altitude” – based on 2- or more-dimensional distances between document coordinates – is estimated by QC. The result is a three-dimensional potential map. The task is to find the optimal landscape in which terms and their documents inhabit their respective contour zones. The procedure is as follows: QC represents documents and terms by Gaussian wave functions whose sum is ψ(x). This means that ψ is modelled as a Parzen window estimator of the form:

$$\psi(x) = \sum_j e^{-\vee x - r_j \vee^2 / 2\sigma^2}$$

By using the Schrödinger equation from quantum mechanics, i.e.:

$$H\psi \equiv \left(\frac{-\sigma^2}{2}\nabla^2 + V(x)\right)\psi = E\psi$$

where *V*(x) is the potential and *E* is the eigenvalue of ψ, we search for the Schrödinger potential for which ψ(x) is the ground state. The minima of the potential function define our cluster centers. In a supervised learning situation, if the number of classes is known beforehand, QC can be fine-tuned for this number and reproduce the original classification by automatic means. The potential *V*(x) can be derived from the previous equation to the following expression:

$$V(x) = E - \frac{d}{2} + \frac{1}{2\sigma^2\psi}\sum_j \vee x - r_j \vee^2 e^{-\vee x - r_j \vee^2 / 2\sigma^2}$$

For a proof, please consult the appendix of [Daránvi & Eklund, 2007].

We can also call in econophysics for a simile. There, the exchange value of shares, stocks, money is fluctuating; however, these are symbols of wealth, not wealth itself. Likewise, we envisage a situation where words, symbols of concepts, have fluctuating meanings dependent on their changing statistical context. Due to this, the actual meaning of a token at time t_n depends on its distance from

the centroid of its lexical field as reproduced by ESOM and evaluated by semantic consistency. Below we translate a financial model example from [Khrennikov, 2010: 155-165] to an ES scenario so that where he discusses financial phase space, we replace it by semantic phase space, and so on.

Let us consider a mathematical model in which a huge number of index terms in a database (the equivalent of a financial market) interact with one another and take into account external social conditions in order to determine the meaning (originally the price) of tokens. We focus on communication as an exchange of words used for indexing, replacing trade with shares of some corporations (e .g. Volvo, Saab, Ikea, etc.).

We consider a *semantic system of coordinates*. We enumerate words with occurrence rates (quasi shares of corporations with market penetration) in the forum of exchange, a term-document matrix: j = 1, 2, …, n (e.g., dog =: j = 1, cat: j = 2, mouse: j = 3, . ..). Introduce the n-dimensional configuration space $Q = R^n$ of word meanings, $q = (q_1, \dots, q_n)$, where q_j is the meaning of an occurrence of the jth word in the indexing vocabulary. Here R is the real line. The dynamics of word meanings is described by the trajectory $q(t) = (q1(t), \dots, q_n(t))$ in the configuration price space Q.

Another variable under consideration is the semantic change variable:

$$v_j(t) = \acute{q}_j(t) = \lim_{\Delta t \to 0} \frac{q_j(t + \Delta t) - q_j(t)}{\Delta t}$$

See for example [Mantegna & Stanley, 2000] on the role of the price change description in comparison. In real models we consider the discrete time scale Δt, $2\Delta t$, … . Here we should use a discrete semantic change variable $\Delta q_j(t) = q_j(t + \Delta t) - q_j(t)$.

We denote the space of semantic changes (drift velocities of terms) by the symbol with coordinates $v = (v_1, \dots, v_n)$. As in classical physics, it is useful to introduce the phase space $Q \times V = R^{2n}$, namely the *semantic phase space*. A pair (q, v) (term meaning, change in term meaning) is called the *state of the "semantic market"*, our semantic space.

Later we shall consider different QL states of the semantic space. The state (q, v) that we consider at the moment is a classical state.

We now introduce an analogue m of mass as the number of term occurrences (similar to shares a trader possesses) that a document "brought" to the market, m being a real number here. We call in the *semantic mass* of a term. Thus each document j (e.g., Volvo in the financial example) has its own financial mass m_j (the rate of respective term occurrences in it). The total semantic content in terms of occurrence rates for the jth document is equal to $T_j = m_j q_j$. Of course this depends on time: $T_j = m_j q_j(t)$. To simplify considerations for now, we model a market in which any term occurrence rates are constant in documents, so m_j does not depend on time. But in principle, our model can be generalized to describe a series of states with time-dependent semantic masses too, i.e. $m_j = m_j(t)$.

We also introduce the *semantic energy* of the state space as a function $H : Q \times V \rightarrow R$. Let us use the analogy with classical mechanics. In this case we could consider (at least for mathematical modeling) the semantic energy of the form:

$$H(q, v) = \frac{1}{2} \sum_{j=1}^{n} m_j v_j^2 + V(q_1, \dots q_n).$$

Here $K(q, v) = ½ \sum_{j=1}^{n} m_j v_j^2$ is the *kinetic semantic energy* and $V(q_1, \dots q_n)$ is the *potential semantic energy*; m_j is the semantic mass of the jth document. The fact that Khrennikov uses term frequencies as term mass whereas we favour PageRank does not change the procedure.

The kinetic semantic energy represents efforts of documents in state space to change term meanings: higher term drift rates induce higher kinetic semantic energies. If the document j_1 has higher semantic mass than document j_2, so $m_{j1} > m_{j2}$, then the same change of meaning, i.e., the same semantic velocity $v_{j1} = v_{j2}$, is characterized by higher kinetic semantic energy: $K_{j1} > K_{j2}$. We also

remark that high kinetic semantic energy characterizes rapid changes in term meanings in the state space. However, the kinetic semantic energy does not give the sign of these changes. It could indicate rapid improvement or the worsening of the semantic consistency of term groups with related meaning.

The potential semantic energy *V* describes the interactions between documents *j* = 1, …, *n* (e.g., competition between documents about cats vs. dogs) as well as external social conditions (e.g., the prevailing topics in the media). For example, we can consider the simplest interaction potential:

$$V(q_1, \dots, q_n) = \sum_{j=1}^{n} (q_i - q_j)^2$$

We could never take into account all social and other conditions that may influence a semantic space. Therefore by using some concrete potential *V*(*t*, *q*) we consider a very idealized model of semantic processes. However, such an approach is standard for physical modeling, where we also consider idealized mathematical models of real physical processes.

Next we apply Hamiltonian dynamics on the semantic phase space. As in classical mechanics for material objects, we introduce a new variable *p* = *mv*, the *semantic momentum* variable. Instead of the semantic change vector *v* = (v_1, … , v_n), we consider the semantic momentum vector *p* = (p_1, … , p_n), $p_j = m_j v_j$. The space of semantic momenta is denoted by the symbol *P*. The space *Ω* = *Q* x *P* will also be called the semantic phase space. *Hamiltonian equations* of motion on the semantic phase space have the form:

$$\acute{q} = \frac{\partial H}{\partial p_j}, \acute{p}_j = \frac{-\partial H}{\partial q_j}, j = 1, \dots, n.$$

If the semantic energy has the form of the Hamiltonian defined above, then the Hamiltonian equations have the form:

$$\acute{q}_j = \frac{p_j}{m_j} = v_j, \acute{p}_j = \frac{-\partial V}{\partial q_j}$$

The latter equation can be written as:

$$m_j \acute{v}_j = \frac{-\partial V}{\partial q_j}.$$

It is natural to call the quantity:

$$\acute{v}_j = \frac{\lim_{\Delta t \to 0} v_j(t + \Delta t) - v_j(t)}{\Delta t}$$

semantic acceleration (rate of change of semantic velocity). The quantity:

$$f_j(q) = \frac{-\partial V}{\partial q_j}$$

is called the (potential) semantic force. We get thereby again the semantic variant of Newton's second law:

$$m\acute{v} = f$$

In other words, the product of semantic mass and semantic acceleration is equal to the semantic force.

We need not restrict our considerations to semantic energies as defined by *H*(*q*, *v*) above. First of all external (e.g. social) conditions as well as the character of interactions between documents in semantic space depend strongly on time. This must be taken into account by considering time-dependent potentials: *V* = *V*(*t*, *q*). Therefore, it can be useful to consider potentials that depend not

only on current semantic values a.k.a. meaning, but also on semantic changes: $V = V(t, q, v)$, or in the Hamiltonian framework: $V = V(t, q, p)$. In such a case the semantic force is not potential. Therefore, it is also useful to consider the semantic version of Newton's second law for general semantic forces as $m\acute{v} = f(t, q, p)$.

From the above it follows that next we can look at semantic pilot waves, a characteristic component of Bohmian mechanics, itself an interpretation variant of QM[25]. In addition to a wave function on the space of all possible configurations, it also postulates an actual configuration of content that exists even when unobserved. The evolution over time of that configuration (that is, of the positions of all particles or the configuration of all fields) is defined by the wave function via a guiding equation. The evolution of the wave function over time is given by Schrödinger's equation.

If, as Khrennikov suggests, we interpret the pilot wave as a field, then it differs crucially from the electromagnetic field. In particular, the force induced by this pilot wave field does not depend on the amplitude of the wave. Thus small waves and large waves disturb the trajectory of an elementary particle to the same extent. Such features of the pilot wave make it possible to speculate [Bohm & Hiley, 1993] that this is just a wave of information (active information). Hence, the pilot wave field describes the propagation of information. The pilot wave is more similar to a radio signal that guides a ship. Of course, this is just an analogy (because a radio signal is related to an ordinary physical held, namely, the electromagnetic field). A more precise analogy is to compare the pilot wave with information contained in the radio signal.

Our fundamental assumption is that documents in semantic space do not display fully classical behavior. Their interactions are ruled not only by classical-like semantic potentials $(t, q_1, \dots, q_n)$, but also (in the same way as in the pilot wave theory for quantum systems) by an additional information potential induced by a semantic pilot wave.

Therefore we cannot use classical semantic dynamics (Hamiltonian formalism) on the semantic phase space to describe the real trajectories of changes in meaning. Information perturbations of Hamiltonian equations for word meaning and changes in it must be taken into account. To describe such a model mathematically, it is convenient to use an object such as a *semantic pilot wave* that rules evolving semantics.

In some sense $\psi(t, q)$, the wave function of the system describes in probabilistic terms the influence of the meaning configuration q on the behavior of documents. In particular, $\psi(t, q)$ contains the expectations regarding documents.

We finish this section by pointing out an important feature of the semantic pilot wave model: all vector space positions charged with meaning are coupled on the information level. The general formalism of the pilot wave theory says that if the function is not factorized (i.e. is entangled),

$$\psi(t, q_1, \dots, q_n) \neq \psi_1(t, q_1) \dots \psi_n(t, q_n),$$

then any drift-related change in the meaning of term q_i will automatically change the behavior of all terms in the system (even those who have no direct coupling with q_i, a familiar assumption from latent semantic models, e.g. [Deerwester et al., 1990]). This will imply a change in the meanings of the rest of the terms j for $\neq \boldsymbol{j}$. At the same time the "hard" semantic potential $V(q_1, \dots, q_n)$ need not contain any interaction term.

For example, let us consider for the moment the potential $V(q_1, \dots, q_n) = q_1^2 + \dots + q_n^2$. The Hamiltonian equations for this potential – in the absence of the semantic pilot wave – have the form $\acute{q}_j = p_j, \acute{p}_j = -2q_j, \ j = 1,2, \dots, n$. Thus the classical semantic value trajectory $q_j(t)$ does not depend on the dynamics in the meaning of other terms $i \neq j$ (for example the fact that the index term "cat"

[25] http://plato.stanford.edu/entries/qm-bohm/

drifted closer to the lexical field of "wild animals" does not depend on the parallel drift in the meaning of "dog", and vice versa). However, if, for example, the wave function has the form:

$$\psi(q_1, \dots, q_n) = ce^{i(q_1q_2+\cdots+q_{n-1}q_n)}e^{-(q_1^2+\cdots+q_n^2)}$$

where $c \in C$ is some normalization constant, then the semantic behavior of documents is entangled, as pointed out in Section 4.2.2.

In summing up the above, we stress again that Bohmian mechanics is peculiar because in spite of QM being probabilistic, it models both the exact position and momentum of a particle as a function of the pilot wave, i.e. the wave function. This, just like the Ehrenfest theorem, creates a kind of particle-wave duality – what used to be a wave of content with a probabilistic location becomes exactly located, i.e. acquires a particle nature, another hallmark of evolving semantics behaving in QL ways.

There exists a parallel between the mapping of scalable evolving semantic content and the Sloan Digital Sky Survey[26]. The latter is concerned with the connection between "dark matter/dark energy" and the geometry of space[27]. The overlap is that the EP extracted from the evolving semantics of the Tate collection represents fluctuating "antigravitational" influence from outside of the system, i.e. social factors counteracting agglomeration tendencies similar to "dark matter", pulling content apart and preventing a content configuration from collapsing onto itself by gravitation.

4.3.3. Index Term Drifts and Entanglement: Integrating Two Analytical Approaches

In what follows we add a software integration feasibility check, testing how Somoclu and Ncpol2spda can work together. This is a potentially interesting question as we will see from the results, but to frame them as evidence for a thoroughly scientific experiment would be wrong for several reasons. Rather, we consider this section as an exercise in hypothesis generation, falling short of hypothesis testing, but still an important step to explore a set of new opportunities brought about by a new toolkit for LTDP. Fig. 4-7 displays the integrated workflow.

We considered the following research question: given that the methodologies developed in PERICLES for studying drifts and quantum-like behaviour integrate and thus provide insights in the nature of evolving semantics as well as on the possible sources of changes or dynamics, what kind of results shall one expect, and what could their interpretation be?

The core idea was to study correlations among distributional patterns that evolve over time. This is similar to the scenario described in Section 4.2.2: there states in a graph structure changed over time, and we studied their correlations to exclude a local hidden variable model. The possible causation between states was restricted by the edges of the graph. Now if we turn our attention to the drift study presented in Section 4.3.1, there any two terms could attract each other, and the magnitude of attraction was based on the mass of each term and their distance. This can also be interpreted as causation that may be present between any two terms.

If we combine the approaches in Sections 4.2.2 and 4.3.1, our task is to detect the temporal correlations between arbitrary pairs of terms in an evolving corpus. The structure of the problem is akin to a fully connected (complete) graph. To study the correlations between nodes in the graph, those nodes also must be assigned a state. In Section 4.2.2, we identified a reference point to anchor the produced emergent self-organizing maps. We can study the motion of the terms relative to this anchor. If over epochs the motion of a term moves toward the centre, we can identify this as a state

[26] [Alam et al., 2016] The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological analysis of the DR12 galaxy sample. Available at http://arxiv.org/pdf/1607.03155v1.pdf
[27] https://www.sdss3.org/surveys/boss.php

+1, and, the other way around, should it move away from the anchor term in the origin, we can interpret it as a state -1. With this, we have all the components for a feasibility study.

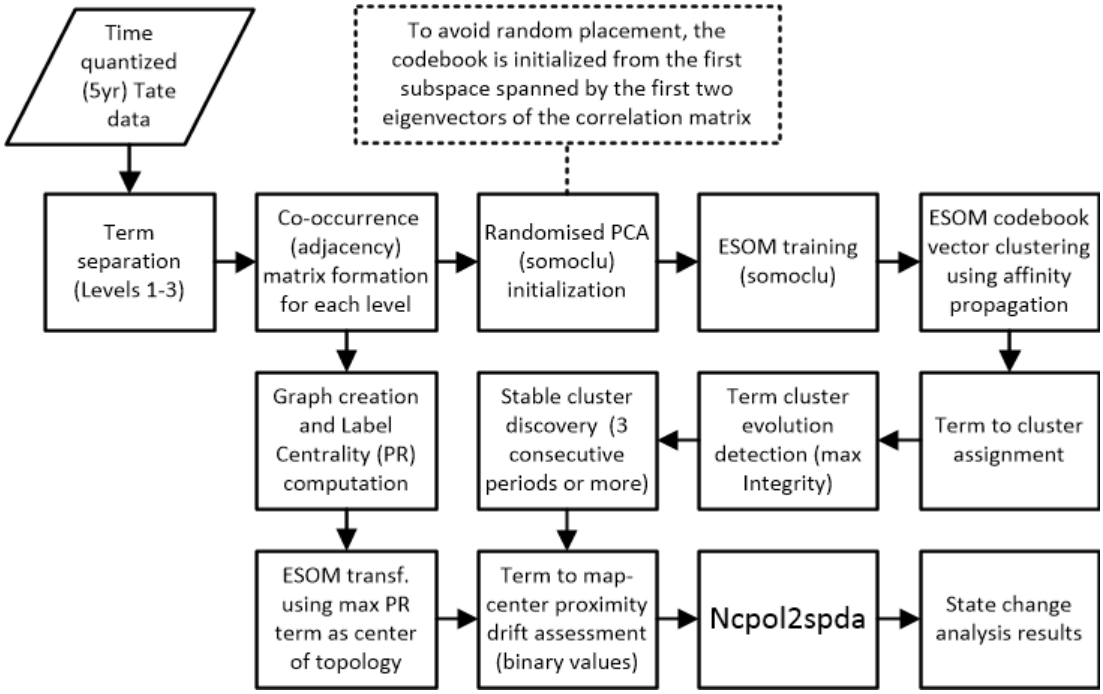

Fig. 4-7. Steps of the integrated workflow for entanglement detection in drift data.

In order to detect temporal correlations we need sets of stable and consistent clusters of terms which we acquire by applying the affinity propagation method on the codebook vectors as mentioned in the framework description in 4.3.1. Specifically, we detect clusters of terms which do not change over time (integrity) and are stable over a window of at least three periods. Stability is defined as the ability of a cluster to appear in as many consecutive periods as possible, whereas to determine the integrity of clusters between timeslots, the Jaccard coefficient is employed [Jaccard, 1912]. The integrity between a pair of consecutive clusters C_{in} and $C_{i(n-1)}$ which appear in periods *n* and *n-1*, is calculated by use of the following formula:

$$J(C_{in}, C_{i(n-1)}) = \frac{|C_{in} \cap C_{i(n-1)}|}{|C_{in} \cup C_{i(n-1)}|}$$

If the similarity is 1, the pair is matched and C_{in} is added to the cluster's timeline. When at least two matches are made (3 clusters in total) the series is added to the temporal correlation test set.

For each of the terms in the selected consecutive clusters, we measure the Euclidean distance between their location and the reference point (term of greatest PR value as mentioned in Section 4.2.2). When the distance becomes smaller over a specific period, then the term has moven towards the reference point and assumes a +1 state. Accordingly, a state of -1 is obviously assumed when the distance becomes larger.

To test this integrated methodology, we returned to the Tate corpus. Within a period, we applied a three-epoch window with drift data by Somoclu in the form of stable term clusters, and studied their changes of states and their correlations by Ncpol2spda. Our results are summarized in the following table (Table 4-5).

Table 4-5. Hidden variable models of drift data indicate feasibility of software tool integration.

Mapsize	Epoch	Level	Local hidden variable model
20x12	1800s	1	Cannot be excluded
40x24	1800s	2	Cannot be excluded
40x24	2000s	2	Cannot be excluded
50x30	2000s	3	Cannot be excluded
60x40	1800s	All	Rejected
60x40	2000s	All	Cannot be excluded

If we cannot exclude a local hidden variable model, this implies that we cannot claim quantum-like behaviour. On the other hand, if we can reject such a model, we can confidently claim the presence of nonlocal correlations as per the predefined causal structure, and hence prove quantum-likeness. The results show that observation granularity, i.e. zooming in on index term specificity, plays an important role, and we were able to detect quantum likeness with all levels included in the period of the 1800s.

The reason why one must be cautious here is that, as Section 4.1 already explained, sentences modeled on QT consider them as content words in entanglement. But given that content words can be defined by ontologies like WordNet [Fellbaum, 1998], for every index term with a definition, a certain measure of entanglement would be imported to the semantic space in analysis, so its traces should be omnipresent. However, only the case of 1800s at 60x40 resolution tested positive, a result in need of a better understanding.

Still, as far as the software integration feasibility check is concerned, the workflow test showed that Somoclu and Ncpol2spda can be used in tandem for new kinds of LTDP-oriented advanced access experiments.

4.3.4. Outlines of an Energy Regime for Word and Sentence Semantics

As we have indicated in Section 3.2, we have found supporting evidence for the QL nature of digital content: although it does not behave completely as described by QM because there is only partial overlap between semantic and QM criteria, neither does it fully comply with CM because it uses a dynamic, relative mass concept instead of specific, constant values to characterize semantic features. As a consequence to indications in 4.2.1-2 and findings in sections 4.3.1-2, below we conjecture that one can augment the concept of a linguistic sign as the unity of form and substance by a third component, its *structuration capacity*. This capacity denotes the "energy" a.k.a. work content inherent in a semantic unit (such as an index term or a machine learning feature), representing amounts of content investment in the construction vs. reconfiguration of semantic spaces, based on the analogy with chemical bonding and compound formation[28].

The basic observation we depart from was made by Salton who suggested that in a dynamic library structured by recursive term vs. document clustering for updates, the cluster centroids keep on being dislocated [Salton, 1975]. This is the earliest reference to semantic drift we know of. Its implication is that index term and respective document dislocations over time are proportional to the quantity and quality of the terms inherent in the update. Whereas this observation has not been followed up as far as we know, it makes perfect sense in our conceptual framework to assume that lexical fields

[28] https://en.wikipedia.org/wiki/Standard_Gibbs_free_energy_of_formation

store context-dependent and relative amounts of work content that can be released from them by interacting with them, enabling bookkeeping-type measurements of semantic content for LTDP. For the quantification of knowledge-related work, see [Ramirez & Steudel, 2008]; for a similar effort to quantify the cognitive extent of science, see [Milojevic, 2015].

In order to explore this idea, one would need an indexing vocabulary represented as a potential surface, in accord with [Taira et al., 2007] or the hypersurface used to model reaction paths in computational chemistry [Mezey, 1999; Hirsch & Quapp, 2004]. As currently no such measurements of term- vs. system state-specific structuration capacities are available, we mention the following arguments in favour of our conjecture:

- Feature content expressed as mathematical energy has been used in ML for quite some time [e.g. LeCun et al., 2008];
- The current energy type we employ to express our model is gravitational (with anti-gravitational hints, see the EP finding in section 4.3.1), falling short of the "dipole" type underlying QM (i.e. EM and spin). Thus terms have relative "masses" fluctuating over time, subject to social pressures, thereby influencing the potential surface whose gradient is the attractive force that manifests comparison between features or objects. However, with energy now a part of the conceptual frame, the Hamiltonian of any system state can be both interpreted and computed;
- The interpretation of the Hamiltonian goes back to the concept of energy in a conservative vector field calculated by the line integral of a moving particle. By analogy, the sum of term drifts can be conceptualized as the symbolic work carried out and stored by a field by taking the sum of respective line integrals. Likewise, given a graph or its corresponding adjacency matrix, the work equivalent of a logical statement or a sentence can be expressed;
- "Term mass" also makes sense in a particle-wave scenario as well where it characterizes wave packets that "behave" as particles (cf. Bohm, see section ...), the scenario being halfway between gravity and QM. Moreover one can express such wave packets as wavelength and respective energy too, as long as one keeps their gravitational origins in mind;
- By this conjecture, we can modify our working definition of QL as the symptomatic coordinated behavior of system state components, described by similar equations in QM and ES.

4.4. Chapter Summary

In this chapter, we demonstrated by a series of experiments that both classical and quantum mechanics offer concepts and scenarios to explore the quickly changing nature of evolving semantic content and information seeking user behaviour, so that our results contribute to mainstream state-of-the-art efforts. The fundamental research problem being knowledge dynamics inherent in scalable distributed digital collections, LTDP must catalog types of changes affecting advanced access to semantic content into the future, and to this end must develop both a conceptual framework and software tools. We took early steps in this direction which can be integrated into ongoing science trends.

We emphasize three of our findings for a short summary:

- In the intersection of classical and quantum mechanics, expressed by the concept of a conservative field with inherent energy (i.e. work content) and described by its Hamiltonian changing with system states, our observations suggest that semantic content inherent in index terms can be modelled on a relaxed, context-dependent version of Newtonian mechanics tentatively called social mechanics. This model will be suitable for scalable studies of both word- and phrase-based indexing of digital objects, and thereby for the synergetic interaction of the LRM and its domain ontologies with computational linguistics in general, bridging the gap between ontological and statistical analyses of evolving content;

- Such methodological studies of knowledge dynamics naturally feed forward to the concept of the Semantic Web both as a distributed knowledge repository and a treasure house for semantic reasoning, posing new challenges for LTDP;
- Mapping the evolving semantic content of scalable collections by physics as a metaphor suggests a parallel with the evolving physical content of the observable universe.

5. Semantic Reasoning for Contextual Content Interpretation

The ability to derive meaningful interpretations from semantic models representing contextualized content and their subsequent evaluation, assessment and storing is of utmost importance to the LTDP domain. In this context, the current chapter discusses our second line of research on contextualized content interpretation based on logical semantics and involving the use of semantic reasoning techniques. After a brief introduction to the background of Description Logics (i.e. the formalism underlying OWL ontologies) and the respective semantic reasoning services, the chapter presents the PERICLES semantic interpretation framework. The latter capitalizes on our adopted representations for contextualized content semantics introduced in D4.4 and is composed of three areas of contribution: (a) ontological inference, namely, deriving implicit knowledge from asserted facts with the use of a reasoning engine, like e.g. Pellet and HermiT; (b) rule-based reasoning, which is a more advanced reasoning approach that is based on rules - in our implementations we are deploying SPIN (SPARQL Inferencing Notation) rules that were originally introduced in D4.4; (c) contextualized reasoning on semantic drifts, which offers the capability of determining the "volatile" and conflicting concepts in an ontology model. Finally, the chapter also presents our proposed scheme for uncertainty management in contextualized content representations, based on non-monotonic and defeasible logics.

5.1. Background

Ontologies are models used to capture knowledge about some domain of interest and their expressiveness and level of formalisation depend on the underlying knowledge representation language used. The **Web Ontology Language** (**OWL**) [Bechhofer, 2009] has emerged as the official W3C recommendation for creating and sharing ontologies on the Web. OWL semantics are based on Description Logics, which are presented in the following subsection, followed by a description of DL-based semantic reasoning.

5.1.1. Description Logics

Description Logics (**DL**) are a family of knowledge representation formalisms characterised by logically grounded semantics and well-defined reasoning services [Baader, 2003]. The main building blocks are concepts representing sets of objects (e.g. `Person`), roles representing relationships between objects (e.g. `worksIn)`, and individuals representing specific objects (e.g. `Alice)`. Starting from atomic concepts, such as `Person`, arbitrary complex concepts can be described through a rich set of constructors that define the conditions on concept membership. For example, the concept $\exists$`hasFriend.Person` describes those objects that are related through the `hasFriend` role with an object from the concept `Person`; intuitively, this corresponds to all those individuals that are friends with at least one person.

A DL knowledge base typically consists of a TBox *T* (terminological knowledge) and an ABox *A* (assertional knowledge). The TBox contains axioms that capture the possible ways in which objects of a domain can be associated. For example, the TBox axiom `Dog` $\sqsubseteq$ `Animal` asserts that all objects that belong to the concept `Dog`, are members of the concept `Animal`, too. The ABox contains axioms that describe the real world entities through concept and role assertions. For example, `Dog(Jack)` and `isLocated(Jack,kitchen)` express that `Jack` is a dog and he is located in the kitchen. Table 5-1 summarises the set of terminological and assertional axioms.

Table 5-1. Terminological and assertional axioms.

Name	Syntax	Semantics
Concept inclusion	$C \sqsubseteq D$	$C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$
Concept equality	$C \equiv D$	$C^{\mathcal{I}} = D^{\mathcal{I}}$
Role Equality	$R \equiv S$	$R^{\mathcal{I}} = S^{\mathcal{I}}$
Role inclusion	$R \sqsubseteq S$	$R^{\mathcal{I}} \subseteq S^{\mathcal{I}}$
Concept assertion	$C(\alpha)$	$\alpha^{\mathcal{I}} \in C^{\mathcal{I}}$
Role assertion	$R(\alpha,b)$	$(\alpha^{\mathcal{I}}, b^{\mathcal{I}}) \in R^{\mathcal{I}}$

The semantics of a DL language is formally defined through an interpretation $\mathcal{I}$ that consists of a nonempty set $\Delta^{\mathcal{I}}$ (the domain of interpretation) and an interpretation function $\cdot^{\mathcal{I}}$, which assigns to every atomic concept A a set $A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$ and to every atomic role R a binary relation $R^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$. The interpretation of complex concepts follows inductively. Table 5-2 shows the syntax and semantics of some of the most common DL constructors.

Table 5-2. Examples of concept and role constructors.

Name	Syntax	Semantics
Top	$\top$	$\Delta^{\mathcal{I}}$
Bottom	$\bot$	$\emptyset$
Intersection	$C \sqcap D$	$C^{\mathcal{I}} \cap D^{\mathcal{I}}$
Union	$C \sqcup D$	$C^{\mathcal{I}} \cup D^{\mathcal{I}}$
Negation	$\neg C$	$\Delta^{\mathcal{I}} \setminus C^{\mathcal{I}}$
Universal Quantification	$\forall R.C$	$\{\alpha \in \Delta^{\mathcal{I}} \mid \forall b.\ (\alpha,b) \in R^{\mathcal{I}} \rightarrow b \in C^{\mathcal{I}}\}$
Existential Quantification	$\exists R.C$	$\{\alpha \in \Delta^{\mathcal{I}} \mid \exists b.\ (\alpha,b) \in R^{\mathcal{I}} \wedge b \in C^{\mathcal{I}}\}$
Inverse	R^{-}	$\{(b,\alpha) \in \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}} \mid (\alpha,b) \in R^{\mathcal{I}}\}$
Transitive Closure	R^{+}	$\bigcup_{n>1} (R^{\mathcal{I}})^{n}$
Composition	$R \circ S$	$R^{\mathcal{I}} \circ S^{\mathcal{I}}$

5.1.2. Semantic Reasoning and DL Reasoning Services

Semantic reasoning (or simply reasoning) is the process of deriving facts and inferring logical consequences from a set of asserted facts or axioms stored in an ontology or knowledge base [Berners-Lee, 1998]. The derived facts are not explicitly stated in the ontology and, thus, constitute the so-called **implicit knowledge** of the ontology. The piece of software capable of performing reasoning is called a **semantic reasoner**, **reasoning engine**, **rules engine**, or simply a **reasoner**. Compared to an inference engine, the reasoner is more generic and provides a richer set of mechanisms to work with.

Besides their formal semantics presented in the previous subsection, DLs come with a set of powerful reasoning services, for which efficient, sound and complete reasoning algorithms with well

understood computational properties are available. Table 5-3 includes the most popular DL reasoners accompanied by short descriptions.

Table 5-3. Existing semantic reasoning technologies.

Reasoner	Description
FaCT++ [Tsarkov & Horrocks, 2006]	Fact++ is a tableaux-based reasoner for expressive Description Logics (DL). It employs a wide range of performance enhancing optimisations, including techniques such as absorption, model merging, ordering heuristics and taxonomic classification.
HermiT [Motik et al., 2009]	Hermit is an OWL reasoning system based on a novel hypertableau calculus. This calculus addresses performance problems due to nondeterminism and model size. HermiT also incorporates a number of other optimizations and techniques towards handling nominals and performing ontology classification more efficiently [Shearer et al., 2008].
Pellet [Sirin et al., 2007]	Pellet is an open-source, Java-based, OWL-DL reasoner with extensive support for reasoning with individuals (including nominal support and conjunctive query), user-defined datatypes, and debugging support for ontologies. It implements several extensions to OWL-DL including a combination formalism for OWL-DL ontologies, a non-monotonic operator, and preliminary support for OWL/Rule hybrid reasoning.
QuOnto [Acciarri et al., 2005]	QuOnto is a Java-based reasoner for the description logic DL-lite with GCIs, which is a sublanguage of OWL DL that can be treated with very efficient database techniques. Reasoning in DL-Lite means not only computing subsumption between concepts, and checking satisfiability of the whole knowledge base, but also answering complex queries.
SHER [Dolby et al., 2009]	SHER is an OWL reasoner that is designed to perform semantic querying of large *Aboxes* using OWL ontologies. SHER proposes an algorithm based on ontology summarization and combines a traditional in-memory description logic reasoner with a database backed RDF Store to scale reasoning to very large Aboxes.

Since the PERICLES domain ontologies are based on OWL and DL, the PERICLES semantic interpretation framework (presented in the following subsections) is built on-top of the established reasoning engined discussed above.

5.2. PERICLES Semantic Interpretation Framework

This section presents the PERICLES semantic interpretation framework and focuses on three directions: (a) Ontological inference, (b) SPIN-based rule reasoning, and, (c) Contextualized interpretations based on drifts.

5.2.1. Representing Content, Context & Use-context

This subsection briefly recaps our adopted ontology-based schemes for semantically representing content, context and use-context (i.e. context of use) in the A&M domain[29]. For a detailed account of these representations, the reader is pointed to project deliverables D2.3.2 [PERICLES D2.3.2, 2015] and D4.4 [PERICLES D4.4, 2016], as well as to [Kontopoulos et al., 2016; Vion-Dury et al., 2015].

SEMANTICALLY REPRESENTING CONTENT

For the A&M domain we have developed three specific domain-related ontologies: (a) the Digital-Video Artwork (DVA), (b) the Software-Based Artwork (SBA), and (c) the Born-Digital Archives (BDA). The developed ontologies are based on several key challenges defined within each of these subdomains and their aim is not to exhaustively model the respective subdomains, but to model specific DP-related risks that demonstrate an interesting range of DP challenges in the domain of interest. The three A&M subdomains share the following common notions:

- **Abstract** (`lrm:AbstractResource`), **Concrete** (`lrm:ConcreteResource`) and **Aggregated Resources** (`lrm:AggregatedResource`) represent the most high-level distinction between resources existing in the domain of interest. An abstract resource is a concept of an entity that may be implemented (`lrm:realizedAs`) in one or more concrete resources. If the realisation of an entity contains more than one resources, then this is represented via an aggregated resource and the different parts are connected with the aggregated instantiation via the property `lrm:hasPart`.
- **Activity** (`lrm:Activity`) represents a Digital Ecosystem activity that may be executed during a digital item's lifespan. An activity can be defined as a temporal action that affects, changes, targets or refers to an item. The A&M domain ontologies extend the Activity class, in order to model domain-specific activities (like for example *creation*, *acquisition*, *storage*, *access*, *display*, *copy*, *maintenance*, *loan*, *destruction* of a DO).
- **Agent** (`lrm:HumanAgent`, `lrm:SoftwareAgent`) represents the entity that may perform an activity or may bring change to the Digital Ecosystem. Human agents are additionally specialised for the A&M domain into artists, creators, programmers, museum staff etc., and software agents into programs, software libraries, operating systems, etc.
- **Dependency** (`lrm:Dependency`) indicates the association or interaction of two or more resources within the Digital Ecosystem that may further affect the functioning or display or existence of a DO. In the A&M ontologies, in order to model complex relationships between resources within the context of each subdomain, we extend the basic notion of `lrm:Dependency` into:
  - **Hardware dependency**, which specifies the hardware requirements for a resource.
  - **Software dependency**, which indicates the dependency of a resource or activity on a specific software agent.
  - **Data dependency**, which implies the requirement of some knowledge, data or information (e.g. passwords, configuration files, input from web service, etc.).

SEMANTICALLY REPRESENTING CONTEXT AND USE-CONTEXT

We represent context via associations between key classes `lrm:Agent`, `lrm:Activity` and `lrm:Resource`, as shown in Fig. 5-1. Agents are related to activities via property `lrm:executes` and its inverse property `lrm:executedBy`. Additionally, when relating an activity to a resource, the latter can be either (a) the resource that is affected by the activity and it is indicated by object

[29] The Space Science domain ontology is based on a different formalism (Topic Maps) and not OWL and DLs, and is thus not discussed in this chapter.

property `:targetsResource` (inverse of `:targetedByActivity`), or (b) a resource that was used during the activity execution, indicated via object property `lrm:used` (inverse of `lrm:usedBy`).

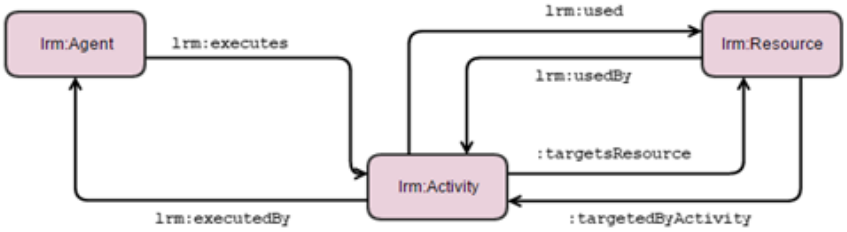

Fig. 5-1. Just Associations between key classes in A&M domain ontologies.

Regarding the representation of use-context, we use `lrm:Dependency` which is explicitly augmented with rich semantics for modelling the underlying preconditions, intentions, specifications and impacts. The notion of *intention* specifies what a dependency intends to express and *specification* thoroughly describes the dependency itself and its context. Furthermore, the notion of *precondition* describes the contextual properties that need to hold in order to consider the dependency as "activated", and the notion of *impact* describes what actions follow when the dependency is activated.

In order to turn dependencies into meaningful correlation links among resources and use-contexts, we have added a set of predefined intention types in order to represent all relevant dependency occasions seamlessly. Below is a description of the proposed intention types [PERICLES D4.3, 2016]:

- **Dependencies with a conceptual intention** are aimed at modelling the intended "meaning" of a resource (i.e. an artwork), according to the way the creator meant for it to be interpreted/understood.
- **Dependencies with a functional intention** represent relations relevant to the consistent and complete operation/functioning of the resource.
- **Dependencies with a compatibility intention** model components which may operate together or as replacement for obsolescence, lack of availability or other reasons.

5.2.2. Ontological Inference

Within the context of ontologies, data is modeled as a set of concepts (classes) that are classified according to a user/domain-defined hierarchy, together with a set of properties and relationships between them. Apart from the explicit declarations of classes, properties and relations, inference techniques can be used to automatically analyze the content of the data and derive facts that are not explicitly expressed in the ontology, revealing thus new relationships and knowledge from the source. "Inference" implies that automated processes can generate new relationships based on the data and on additional information that are given in the form of rules, axioms, relations, constrains, etc. within the ontology.

As already discussed before, a semantic reasoner is a piece of software that can be used to infer logical consequences from a set of asserted facts and declarations. The reasoner takes into account the DL in the ontology in order to perform tasks such as: (a) checking for potential inconsistencies, (b) automatically classifying ontology notions into classes, (c) discovering new knowledge (e.g. assigning property values or interlinking previously independent classes), and (d) executing SPARQL queries [Prud'Hommeaux & Seaborne, 2008]. In our examples, we use Pellet[30], a well-known OWL 2 DL reasoner which can be added as a plugin in the Protege ontology editor. However, any other DL-

[30] https://github.com/Complexible/pellet

enabled reasoner would equally suffice, either as ontology editor/IDE plugins, or as standalone tools that can be used programmatically from within third-party applications. The following subsections report on indicative examples of inferred knowledge that can be derived by a DL reasoner operating on top of the A&M domain ontologies.

Inference Based on Class Axioms

OWL 2 enables the representation of knowledge of a domain by using class expressions and property restrictions. Restrictions should be considered as part of the meaning of a class or property, and thus they participate in the classification process of entities or in the membership definition of individuals in a class.

In the A&M domain, we have defined restrictions in entities, which are expressed as `owl:equivalentTo` axioms according to the principles of *Manchester OWL syntax* [Horridge et al., 2006]. First, there are axioms that allow the reasoner to classify an entity (i.e. instance) as a Digital Video Artwork, or as a Software-Based Artwork, or as a specific type of Born-Digital Archive. More specifically, for the case of `DigitalVideoArtwork`, the axiom is expressed as:

```
dva:DigitalVideoArtwork ≡ (lrm:hasPart some dva:DigitalVideo) and (realizes
some dva:DigitalVideoArt)
```

meaning that an entity that `:hasPart` one or more instances of `:DigitalVideo` and additionally is the realization of an instance of `:DigitalVideoArt`, should be classified as a `:DigitalVideoArtwork`.

In SBA, the axiom that classifies an instance as of type `:SoftwareBasedArtwork` is the following:

```
sba:SoftwareBasedArtwork ≡ (lrm:hasPart some (lrm:SoftwareAgent or
sba:DigitalResource)) and (realizes some lrm:SoftwareBasedArt)
```

which is interpreted by the reasoner as if there is an entity that `:hasPart` one or more instances that are either of `lrm:SoftwareAgent` or of `sba:DigitalResource` type, and additionally it is the realization of an instance of `:SoftwareBasedArt`, then this entity should be assigned into the `:SoftwareBasedArtwork` class of SBA ontology.

Similarly, in BDA, there are axioms that assign an instance into a specific class of Born Digital Archive, which corresponds to the actual level of description of the archival material (for more details, see ISAD(G) standard in [ICA, 2000]). The key property in these axioms is the `dva:hasLevelOfDescription`, as seen below:

```
bda:Fonds ≡ (dva:hasLevelOfDescription value "Fonds") or
(dva:hasLevelOfDescription value "Sub-fonds")
bda:Series ≡ (dva:hasLevelOfDescription value "Series") or
(dva:hasLevelOfDescription value "Sub-series")
bda:File ≡ dva:hasLevelOfDescription value "File")
bda:Item ≡ dva:hasLevelOfDescription value "Item")
```

Based on the aforementioned declarations, the reasoner may infer that an entity is of `bda:Fonds` type if the value of its `:hasLevelOfDescription` property is either "`Fonds`" or "`Sub-fonds`". Similar interpretations can be made for the `Series`, `File` and `Item` classes.

Moreover, as already seen in D4.4 [PERICLES D4.4, 2016], for the case of `Dependency` and for all three subdomains we have defined that:

```
:HardwareDependency ≡ lrm:Dependency and (lrm:from some :Equipment)
:SoftwareDependency ≡ lrm:Dependency and (lrm:from some lrm:SoftwareAgent)
```

```
:DataDependency ≡ lrm:Dependency and (lrm:from some(lrm:DigitalResource or lrm:Description)
```

By interpreting the above `equivalentTo` axioms, a reasoner may infer new knowledge regarding the classification of a Dependency into a further specialized type:

- If there is an instance of `Dependency` in the ontology that is related to one or more instances of `:Equipment` type by the `lrm:from` property, then it is a `HardwareDependency`.
- If there is an instance of `Dependency` in the ontology that is related to one or more instances of `lrm:SoftwareAgent` type by the `lrm:from` property, then it is a `SoftwareDependency`.
- If there is an instance of `Dependency` in the ontology that is related to one or more instances of either `:DigitalResource`[31] or of `:Description`[32] type, via the `lrm:from` property, then it is a `DataDependency`.

Similar restrictions have been defined for the subclasses of the `:Activity` class of the A&M subdomain ontologies. Instances that are connected with Resources by specific subclasses of `:targetsResource`, can be assigned automatically by the reasoner into corresponding subclasses of `:Activity`. For example, the following axiom:

```
:AccessActivity ≡ :accessesResource some lrm:Resource
```

states that if an entity is linked with one or more Resources via the `:accessesResource` property, then the entity is (or should be) an instance of `:AccessActivity` class. If the instance is already classified by the user as a different and disjoint type of class, then the reasoner will identify this state as an inconsistency.

A more complex axiom is expressed for the `:CopyActivity` where two specific relations should exist in order to classify an instance as of `:CopyActivity` type:

```
CopyActivity ≡ (:hasCopyOutput some lrm:Resource) and (:hasCopyInput exactly 1 lrm:Resource)
```

meaning that in order to classify an instance as of `:CopyActivity`, there should be a relation that connects it with exactly one entity of Resource type via the `:hasCopyInput` property, and also with one or more instances of Resource type via the `:hasCopyOutput` property.

Inference Based on Property Characteristics

Transitive Properties

A transitive property `P` declares that if `P(x,y)` and `P(y,z)` are defined, then `P(x,z)` can be implied. In other words, a transitive property links two individuals `x` and `z`, whenever a connection between `x` and `y` and between `y` and `z` is defined, for some individual `y` and for the same property.

In all A&M subdomain ontologies, we define property `:isCopyOf` as a transitive property; this property is useful for connecting resources under the process of a `:CopyActivity`. The reasoner, acting on top of the ontology, may lead to a new relationship between two unrelated nodes, as can be seen in the triples below (the red text denotes the implicit/inferred knowledge):

```
software_based_artwork_1_copy :isCopyOf software_based_artwork_1
```

[31] An instance of `:DigitalResource` type involved in a `:DataDependency` might represent cases where a resource depends on a specific file (raw data, script, etc.) as input in order to operate properly or to achieve its purpose of existence.
[32] An instance of `:Description` type involved in a `:DataDependency` might represent cases where a resource depends on specific data or knowledge (i.e. for example an encryption key, a password) in order to operate properly or to achieve its purpose of existence.

```
software_based_artwork_2_copy :isCopyOf software_based_artwork_1_copy
software_based_artwork_2_copy :isCopyOf software_based_artwork_1
```

Furthermore, the `lrm:hasPart` property that is used to connect an aggregated resource with its components, is declared as a transitive property. An example of inferred relations can be seen with a red arrow in Fig. 5-2, where new inferences between unrelated nodes, i.e. between (a) realization of Becoming and LCD monitor, (b) realization of Becoming and a computer, are defined by the reasoner, due to the transitive characteristic of the `:hasPart` property.

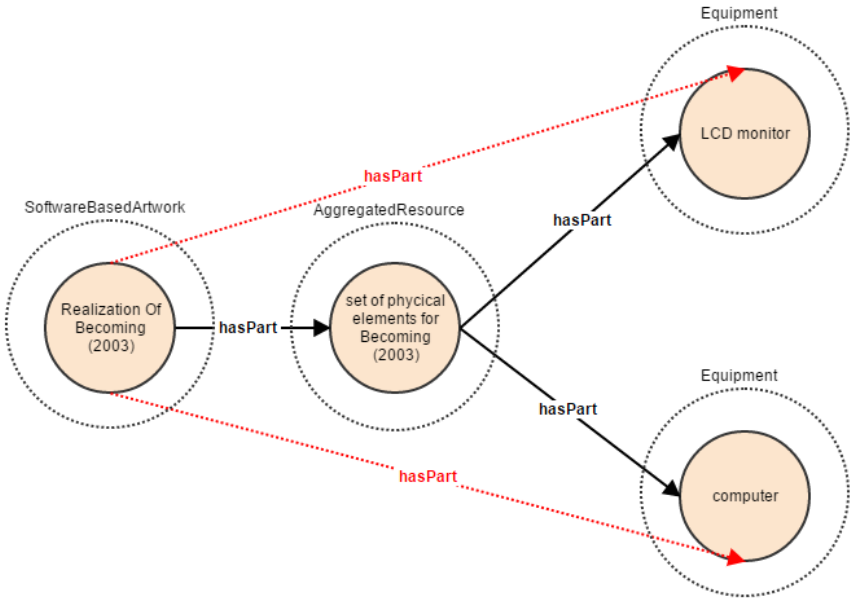

Fig. 5-2. User-defined and reasoner-inferred relations between entities via the `:hasPart` property, due to the transitive character of the property.

Additional properties, adopted from LRM, which are of transitive type are the `lrm:preceding` and its inverse property named `lrm:following`. The aforementioned properties are used to define a strict causal ordering between any `lrm:Resource` instances. An example of asserted and inferred knowledge is given in the following triples:

```
delta_1 :preceding delta_2
delta_2 :preceding delta_3
delta_1 :preceding delta_3
```

which means that the change described via `delta_1` (instance of `lrm:RDF-Delta` type) happened before the change represented via `delta_2`, and also `delta_2` took place before `delta_3`; the obvious conclusion that `delta_1` also happened before `delta_3` becomes feasible due to the transitive characteristic for that property declared in the ontology.

Functional Properties

A functional property indicates that at most one distinct value can be assigned to any given individual via this property. Such declaration leads the reasoner to infer that two different nodes that are connected with the same instance via a functional property, are actually the same; or differently stated, if `P(x,y)` and `P(x,z)` exist, then `y` and `z` should be the same (`y==z`).

A characteristic example of such property could be the `:identification` property (adopted in A&M ontologies from LRM) that relates a digital resource to a unique ID. If two entities have the same ID as declared, the reasoner will state that these two entities refer to the same resource.

Another set of properties that are declared as functional are those defined in our proposed **digital video ontology design pattern** (**ODP**) [Mitzias et al., 2015]; these are the properties that represent the technical characteristics of a digital video in the ontology, like for example: `:hasAspectRatio`, `:hasCodec`, `:hasBitRate`, `:hasFrameRate`, etc. Since properties are of functional type, the reasoner will interpret that two instances that are connected with the same resource via the property are the same, even if the explicit relationships in the ontology might declare otherwise. An example case and the inferred knowledge are shown in the following triples:

```
video_stream_1 :hasAspectRatio 5:4
video_stream_1 :hasAspectRatio five_to_four
5:4 owl:sameAs five_to_four
```

Inverse Functional Properties

This property plays a similar role as the functional property, but the relation between entities is reversed: if `P(y,x)` and `P(z,x)` then it can be inferred that `y` and `z` are the same (`y==z`). In other words, a single value of the property cannot be shared between two entities. If two entities are found to share the same value for an inverse functional property, then these two entities are inferred to be the same.

Again, the `:identification` property is declared in A&M domain ontologies as inverse functional, since the ID attached as a value of the property to a given entity should be unique and could not be assigned as an ID in a different entity. If two different entities have the same ID via the `:identification` property, then they are considered by the reasoner as the same (see Fig. 5-3).

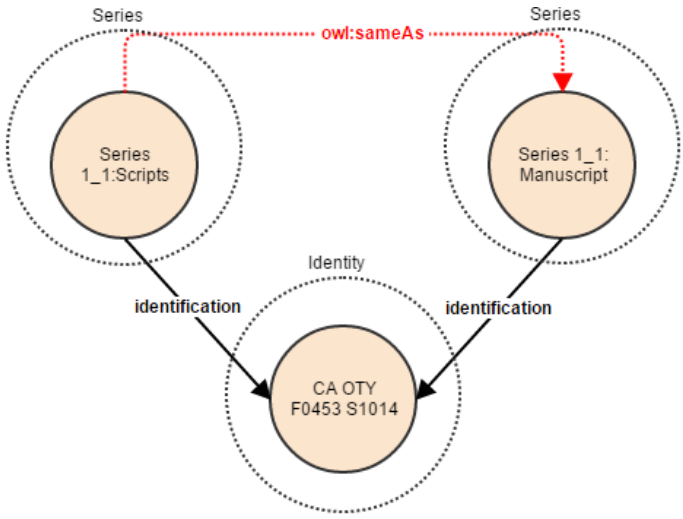

Fig. 5-3. User-defined and reasoner-inferred relations between entities via the `:identification` property, due to the inverse functional character of the property.

Symmetric & Asymmetric Properties

A symmetric property is a property for which holds that if the pair `(x,y)` is an instance of `P`, then the pair `(y,x)` is also an instance of `P`. A symmetric property implies that a relationship is bidirectional even if the relationship was only modeled in one direction. In such case, the reasoner may infer an additional relationship between two related nodes. The case of symmetric properties couldn't be applied to any of the properties in the A&M domain ontologies.

An asymmetric property is the opposite of symmetric, which means that it prevents a symmetrical inference; if an individual `x` is connected by an asymmetric property `P` to an individual `y`, then `y` cannot be connected to `x` via the same property `P`. In a case where a bi-directional relationship is

stated between two entities with the same property then the reasoner will produce an inconsistency error and the user will have to correct the declaration correspondingly.

Asymmetric and symmetric properties are reasonable only in cases where the connection is eligible between resources that are of the same type. In the A&M domain ontologies, the only properties that have declared the same type of class in their `rdf:domain` and `rdfs:range` values, and at the same time they can be of asymmetric type, are the `:isCopyOf` and the `:preceding` and `:following`; these properties also happen to be of transitive type (as mentioned in corresponding section).

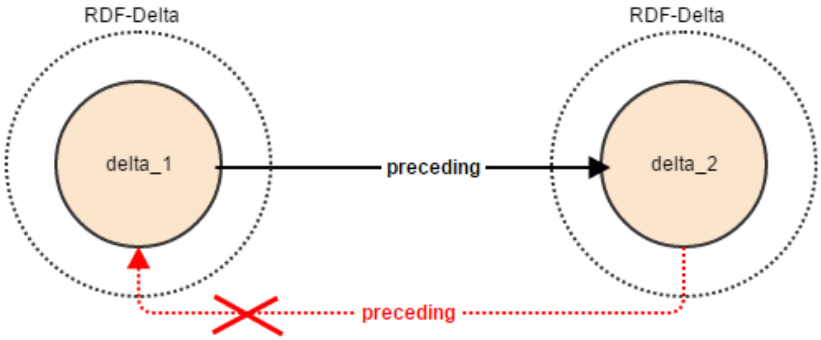

Fig. 5-4. User-defined relation between entities via the `:preceding` property, and relation that is prevented by the reasoner, due to asymmetric character of the property.

Reflexive & Irreflexive Properties

A reflexive property in OWL 2 relates everything to itself. This does not necessarily mean that every two individuals related via a reflexive property are the same. It becomes evident that reflexive properties are reasonable if the property connects entities that are of the same type. In the A&M domain ontologies, there was no need to define any property as reflexive.

An irreflexive property means that the axiom that it describes, prevents a reflexive inference. If an irreflexive axiom contradicts specific instantiations in the ontology, then the ontology will become inconsistent. Similarly as in reflexive properties, the connected entities should be of the same type, otherwise the declaration of a property that has different `rdfs:domain` and `rdfs:range` values is redundant. In the A&M domain ontologies none of the defined properties could act as irreflexive.

Inference based on Domain and Range Restrictions

For every property defined in the A&M domain ontologies, specific restrictions in their domain and range declarations were stated [PERICLES D2.3.2, 2015; Sections 8.2, 8.4 and 8.6], enriching this way the structure, the semantics and the context of the ontologies. These declarations may be further taken into account by the reasoner in order to produce proper inferences regarding the classification of instances.

In more detail, the `rdfs:domain` is an instance of `rdf:Property` that is used to state that any resource that has a given property is an instance of one or more classes, as those are defined in the object of the triple <`P rdfs:domain C1`>. Anything that is related by *P* to something else, must be a *C1*. On the other hand, the `rdfs:range` is an instance of `rdf:Property` that is used to state that the values of a property are instances of one or more classes, as those are defined in the "object" of the triple <`P rdfs:range C2`>. Anything to which something is related by *P* must be a *C2*.

As seen in example properties in Table 5-4, the domain and range strictly define the class of instances that each property can relate. This means that in case where the class-type of an instance that takes part in a property instantiation is not defined, the reasoner will infer a classification result for that

instance, in the form of <instance_1 rdf:type class_type>, where class_type will be consistent to the rdfs:domain or rdfs:range of the property.

Table 5-4. Examples of rdfs:domain and rdfs:range declarations of properties in the A&M subdomain ontologies.

Ontology Name	Object/Data Property	rdfs:domain	rdfs:range
DVA	:hasAspectRatio	:VideoStream (*subClassOf* :Stream)	:AspectRatio (*subClassOf* :VideoDescription)
SBA	:hasSourceCode	:SoftwareAgent (*subClassOf* :Agent) or :DigitalResource (*subClassOf* :ConcreteResource)	:SourceCode (*subClassOf* :DigitalResource)
BDA	:hasLevelOfDescription	:File (*subClassOf* :AggregatedResource) or :Fonds (*subClassOf* :AggregatedResource) or :Series (*subClassOf* :AggregatedResource) or :Item (*subClassOf* :DigitalResource)	{"File","Fonds","Item","Series","Sub-fonds","Sub-series"} (string with one of the aforementioned values)

There are cases, like the :hasSourceCode or :hasLevelOfDescription property, where if an unclassified instance x is the object of the RDF-statement where the property takes part, then x may belong to *any* (union set) of the declared classes in the rdfs:domain of the property, without affecting the consistency of the ontology.

Similarly, there are cases of data properties where the rdfs:range can be either any of the OWL built in datatypes (string, integer, date, etc.) or a data range expression given by the user (e.g. see rdfs:range in :hasLevelOfDescription property). Such declarations are considered by the reasoner as sufficient conditions in order to detect inconsistent cases (like for example if a level of description is not defined as one of the values given as valid in the rdfs:range declaration of the :hasLevelOfDescription property, then the ontology would be inconsistent).

5.2.3. SPIN Reasoning Layer

The second component of the PERICLES semantic interpretation framework builds on the early work presented in D4.4, where a SPIN reasoning layer was presented for detecting content- and context-based inconsistencies in the A&M subdomain ontologies [PERICLES D4.4, 2016]. SPIN, the **SPARQL Inferencing Notation** [Knublauch et al., 2011], is a well-known notation for representing SPARQL rules and constraints on models, and for performing queries on RDF graphs. SPARQL queries can be stored as RDF triples alongside the RDF domain model, enabling the linkage of RDF resources with the associated SPARQL queries, as well as their consequent sharing and reuse. SPIN can also be used to derive new RDF statements from existing ones through iterative rule application.

An alternative to SPIN, which we initially considered instead, is **SWRL** (**Semantic Web Rule Language**) [Horrocks et al., 2004]. However, although both languages have an RDF syntax for representing rules, SPIN is superior to SWRL in almost every respect, mainly because SPIN is based on SPARQL, which is well established and supported by numerous engines and databases. This means that SPIN rules can be directly executed on the databases and no intermediate engines with communication overhead need to be introduced. Furthermore, regarding expressiveness, SPIN is significantly more expressive, because SPARQL has various features such as UNIONs and FILTER expressions. Also, SPIN has an object-oriented model that arguably leads to better maintainable models than SWRL's flat rule lists. Additionally, SPIN goes far beyond being just a rule language, and also provides means to express constraints and to define new functions and templates. Finally, although both languages have reached the same standard status (i.e. W3C Member Submissions), SWRL is no longer actively maintained and its usage is now strongly discouraged. Based on SPIN's advantages outlined above, we have deployed SPIN as the foundation of our rule-based reasoning layer within T4.5.

SPIN can be used from within TopBraid Composer (TBC)[33], a popular ontology editing IDE. Programmatically, one can also use an open source library called TopBraid SPIN API[34], which has intentionally been designed to be independent from any other TopBraid-related dependencies, so that it can be used in any conceivable Java application including servlets. The SPIN API is built on the Apache Jena API[35].

Detection of Inconsistencies with SPIN

In the A&M ontologies, SPIN rules are used for taking advantage of elements from the context of digital resources in order to detect inconsistencies while examining a specific state of the digital ecosystem, or for cases where SPIN rules monitor policies existing in the digital ecosystem in order to trigger changes that policies describe. Examples of both cases are given below; the DP-related risk scenarios for the three A&M subdomains have been adopted from [Falcao, 2010] and [Rice, 2015] and an early implementation was presented in [Lagos et al., 2016]; the policies example expresses precondition and impact of dependencies as SPIN rules and tracks a policy of a real case scenario and performs a change in the ecosystem accordingly.

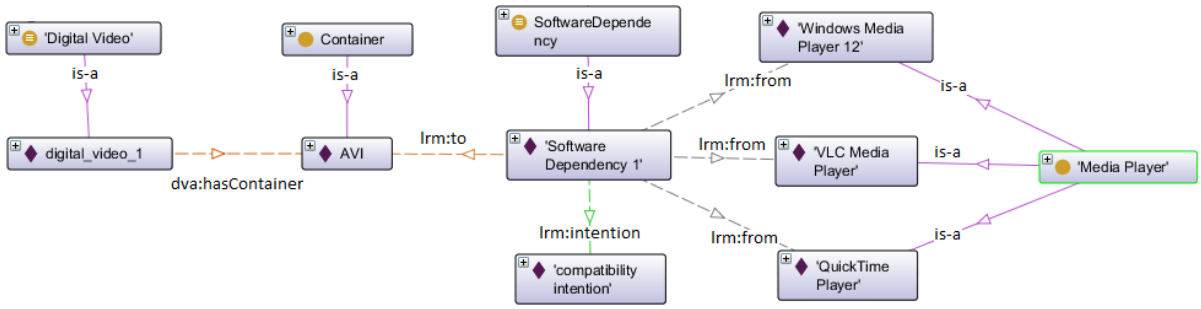

Fig. 5-5 Association of a digital video with a container and the container's software dependency from media players.

DVA Risk Scenarios

Video playback failure due to unsupported container - Digital videos (`dva:DigitalVideo`) are usually associated with containers (`dva:Containers`) with the property `dva:hasContainer`. A container (or wrapper) contains the various components of a video, such as video and audio streams.

[33] http://www.topquadrant.com/tools/modeling-topbraid-composer-standard-edition/
[34] http://topbraid.org/spin/api/
[35] http://jena.apache.org/

Moreover, a software dependency (`dva:SoftwareDependency`) with *compatibility intention* is often used to represent compatibility between containers and specific media players (`dva:MediaPlayer`), as seen in Fig. 5-5.

Since a digital video's playback activity (`dva:PlaybackActivity`) utilizes (`lrm:used`) a media player in order to play the video, it is vital for the success of the activity that the selected player supports the video's container. The satisfaction of this requirement could be examined with the use of the SPIN rules below:

```
CONSTRUCT
{
      ?activity         a                        dva:ErrorItem .
      ?activity         dva:hasErrorText         "Incompatible player for
                                                 playback activity" .
}

WHERE
{
      ?digital_video    a                        dva:DigitalVideo .
      ?digital_video    dva:hasContainer         ?container .

      ?dependency       a                        dva:SoftwareDependency .
      ?dependency       lrm:to                   ?container .

      ?activity         dva:playsResource        ?digital_video .
      ?activity         lrm:used                 ?player .

      MINUS
      {
            ?dependency       lrm:from           ?player .
      } .
}
```

The application of this set of rules would classify, in case of inconsistency, a playback activity as an error item (`dva:ErrorItem`) and present an explicit error text.

Inconsistent video playback due to missing aspect ratio information - In the case that a container's metadata do not carry information on the aspect ratio (`dva:AspectRatio`), a media player might proceed to automatically selecting a default aspect ratio. Such a case could be inconsistent, as it might affect the aesthetic intention of the artwork, and human intervention should be invoked by presenting an appropriate warning. A SPIN rule, based on the object property `dva:includesAspectRatio`, would be:

```
CONSTRUCT
{
      ?digital_video    a                        dva:WarningItem .
      ?digital_video    dva:hasWarningText       "No aspect ratio
                                                 information in container"
.
}

WHERE
{
      ?digital_video    dva:hasContainer         ?container .
      ?digital_video    a                        dva:DigitalVideo .
      ?container        dva:includesAspectRatio  false .
}
```

In this scenario, SPIN rules classify resources as warning items (`dva:WarningItem`), rather than error items, as the video would still be playable, and it relies on the DP expert to judge whether the result is acceptable.

SBA Risk Scenarios

Execution failure due to incompatible operating system - A software-based artwork (`sba:SoftwareBasedArtwork`) usually consists of one or more executables (`sba:ExecutableFile`). Such files are frequently compatible with certain operating systems (`sba:OperatingSystem`), resulting in failure of the display activity (`sba:DisplayActivity`) if the used computer incorporates an incompatible operating system.

The case of a software-based artwork named *Becoming* by Michael Graig-Martin[36] could be used as an example. An earlier realization of the artwork includes an executable which is compatible only with older versions of Microsoft Windows. This information is represented with a dependency (`sba:SoftwareDependency`), as seen in Fig. 5-6.

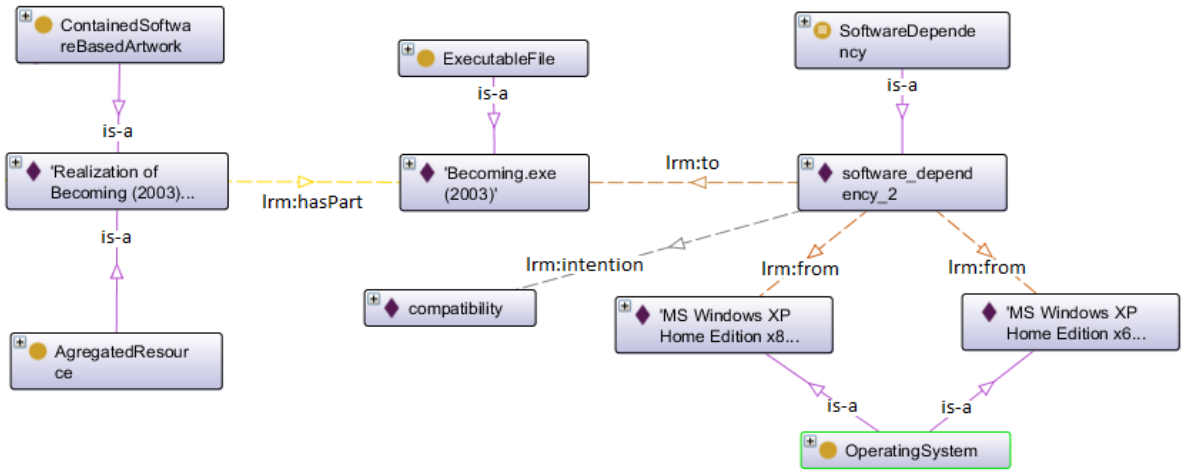

Fig. 5-6. Becoming executable dependency from specific operating systems.

In order to perform a display activity of the artwork, the computer in use should operate with one of the compatible versions of Windows XP. This consistency check can be appointed to the set of SPIN rules that follows:

```
CONSTRUCT
{
      ?activity          a                        sba:ErrorItem .
      ?activity          sba:hasErrorText         "Incompatible operating
                                                  system" .
}

WHERE
{
      ?activity          sba:displaysResource     ?software_based_artwork .
      ?activity          lrm:used                 ?computer .

      ?computer                sba:usesSoftware   ?operating_system .
      ?operating_system        a                  sba:OperatingSystem .
```

[36] http://www.tate.org.uk/art/artworks/craig-martin-becoming-t11812

```
      ?software_based_artwork lrm:hasPart       ?executable .
      ?executable             a                 sba:ExecutableFile .

      ?dependency             lrm:to            ?executable .
      ?dependency             a                 sba:SoftwareDependency .

      MINUS
      {
            ?dependency       lrm:from    ?operating_system .
      } .
}
```

The application of this check would result in classifying the display activity as an error item in case of a computer with incompatible operating system.

Execution failure due to different version of external APIs - Certain SBAs utilize external resources, such as network connections or third-party APIs. For example, the artwork *Brutalism*[37] performs searches to Google via the Google API and prints out the search results. Obviously, the artwork's operability is vulnerable to API updates or changes with no backward compatibility, as they may lead to malfunction or no function at all. In such a case, a curator should be notified by a warning message so as to examine the features of the newer API. The representation of the artwork in the SBA domain ontology, along with a display activity, can be seen in Fig. 5-7.

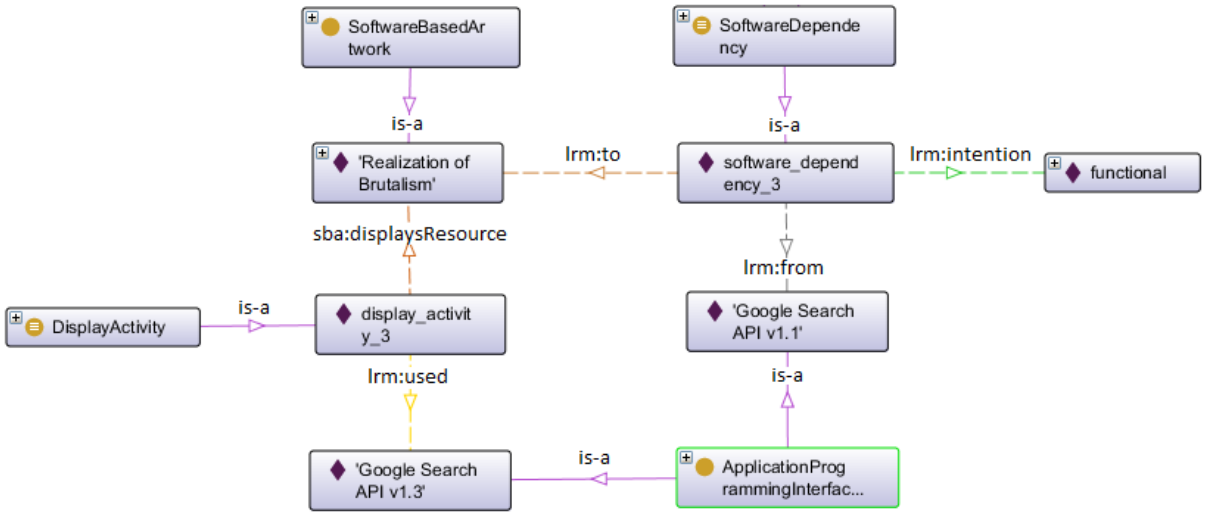

Fig. 5-7 Display activity of Brutalism and dependency from Google API.

In SPIN, the consistency check would be expressed as:

```
CONSTRUCT
{
      ?activity         a                       sba:WarningItem .
      ?activity         sba:hasWarningText      "Varying API version" .
}

WHERE
{
      ?activity         sba:displaysResource    ?software_based_artwork .
```

[37] http://www.tate.org.uk/art/artworks/martinat-mendoza-brutalism-stereo-reality-environment-3-t13251

```
        ?api                a           sba:ApplicationProgrammingInterface .

        ?dependency         a           sba:SoftwareDependency .
        ?dependency         lrm:to      ?software_based_artwork .
        ?dependency         lrm:from    ?api .

        MINUS
        {
              ?activity          lrm:used          ?api .
        } .
}
```

Similarly to the previous example, a display activity using a different API would be classified as a warning item.

BDA Risk Scenarios

Normalization activity failure caused by incompatible normalization software - A digital file, e.g. a text document, may be processed through a normalisation activity (`bda:NormalisationActivity`), using a certain program. It is usually the file format (`bda:FileFormat`) that defines which software should be used, since, in many cases, an incompatible program might either be unable to open the file or it may open it incorrectly, messing the text formatting, fonts, etc. The affinity between a file format and some software is represented as a software dependency (`bda:SoftwareDependency`). Fig. 5-8 shows a normalisation activity representation, along with a software dependency, that defines the compatible programs.

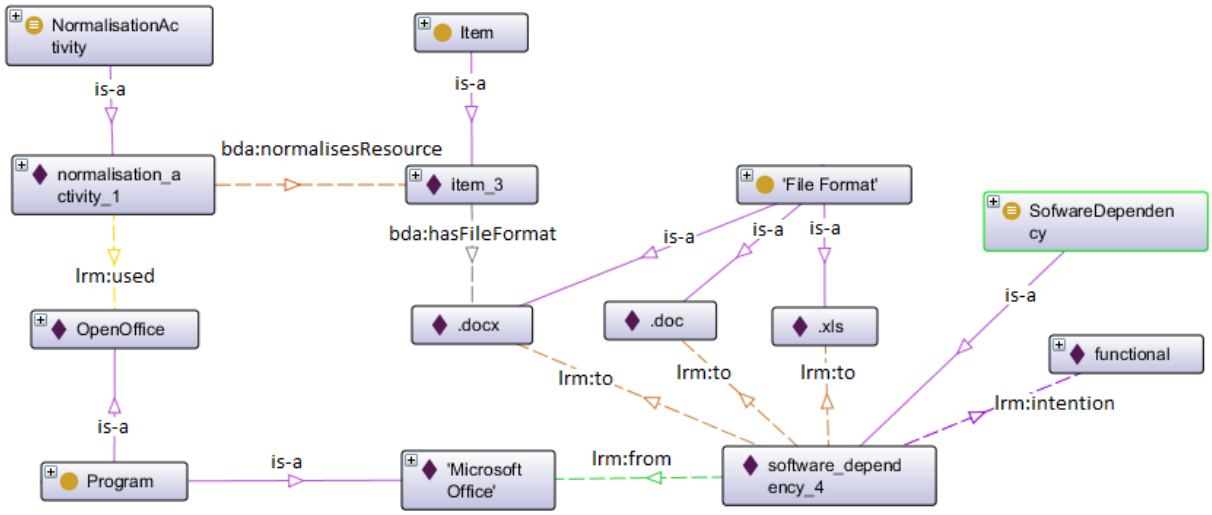

Fig. 5-8. Normalisation activity of a text document, along with a software dependency.

In a nutshell, during the activity shown in Fig. 5-8, OpenOffice was used to normalise a file with a *.doc* extension, while the software dependency indicates that Microsoft Office is the appropriate program when it comes to file formats such as *.docx*, *.doc* and *.xls*. Thus, an error message should inform the operator about this inconsistency, which could be generated by the application of the SPIN rules shown next:

```
CONSTRUCT
{
        ?activity          a                        sba:ErrorItem .
        ?activity          bda:hasErrorText         "Incompatible software" .
```

```
}

WHERE
{
       ?activity          bda:normalisesResource  ?item .

       ?program           a                       bda:Program .

       ?item              bda:hasFileFormat       ?file_format .

       ?dependency        a                       bda:SoftwareDependency .
       ?dependency        lrm:to                  ?file_format .
       ?dependency        lrm:from                ?program .

       MINUS
       {
              ?activity   lrm:used     ?program .
       } .
}
```

SPIN AND POLICIES

In the context of PERICLES, we define **policies** as a set of obligations, which may refer either to the state of things ("*this file was accessed one week ago*") or processes to be performed, and may result in a traceable state (e.g. "*approval for disposal of a document must be givenmade by the approved member of staff*"). Policies are defined in order to support, among others, data management requirements and rule-based change management tasks.

Concrete implementation of policies for change management is highly dependent on the use case. Nevertheless, LRM [PERICLES D3.3, 2015] is used as a common base ontology language for change management. More specifically, the model allows to express change to entities in the ontologies using deltas (`lrm:RDF-Delta`). Deltas provide meta-information about the modification of a resource, by defining a list of triples that have been deleted and added to the model. Moreover, the LRM introduces dependencies (`lrm:Dependency`) that can model here associations between policies and DOs within the digital ecosystem, as well as the concepts of *precondition* and *impact*, as means to handle change in the digital entities. The precondition describes the conditions that have to be satisfied to activate a dependency, while the impact describes the consequences of the dependency activation.

By defining dependencies that make use of these constructs, we propose to implement policies and change management at the model-level, expressed as constraints on entities in the corresponding LRM model. In order to accomplish this type of policy implementation, it is necessary to have support for rule languages at the model level; here we use the W3C SPIN rule standard[38].

For every instance of dependency, and through precondition and impact instantiations, we define specific SPIN rules that are triggered upon a new change (delta). In [PERICLES D5.3, 2016], and in order to present the efficiency of the approach, we describe exemplar implementations of SPIN rules in correlation to specific policies and to specific change scenarios:

- **Change in the value of a policy parameter** (e.g. change in the total eligible time under which a DO can be stored in a private repository, change in threshold drift for concepts of interest, etc.)
- **Change in the value of a DO parameter** (e.g. change in the current total time during which the DO is stored).

[38] https://www.w3.org/Submission/spin-overview/

5.2.4. Using Background Domain Knowledge

Once a domain-specific ontology is created, the next step is to populate the empty schema with background domain knowledge. As such we consider any information that is relevant to the realisations of concepts (i.e. instances) and their relations in the domain, and not to the concept hierarchies and the structure itself.

If done manually, the **ontology population** process, i.e. the instantiation of new knowledge in an ontology-based representation, can be a time-consuming and error-prone task. As a result, research has shifted attention to automating the process of identifying and adding new instances from an external source into an ontology [Buitelaar & Cimiano, 2008].

Within the context of PERICLES, we have developed **PROPheT**[39], a novel application that enables instance extraction and ontology population from Linked Open Data (LOD) sources, such as DBpedia[40] and Europeana[41], through a user-friendly graphical user interface [Mitzias et al., 2016; PERICLES D4.3, 2016]. PROPheT offers access to available LOD sources, facilitating through different types of instance extraction-related functionalities the discovery, reusability and extensibility of knowledge in any domain of interest. PROPheT simplifies the way of communicating information with LOD sources, without needing a high level of expertise for querying, accessing and storing the available data.

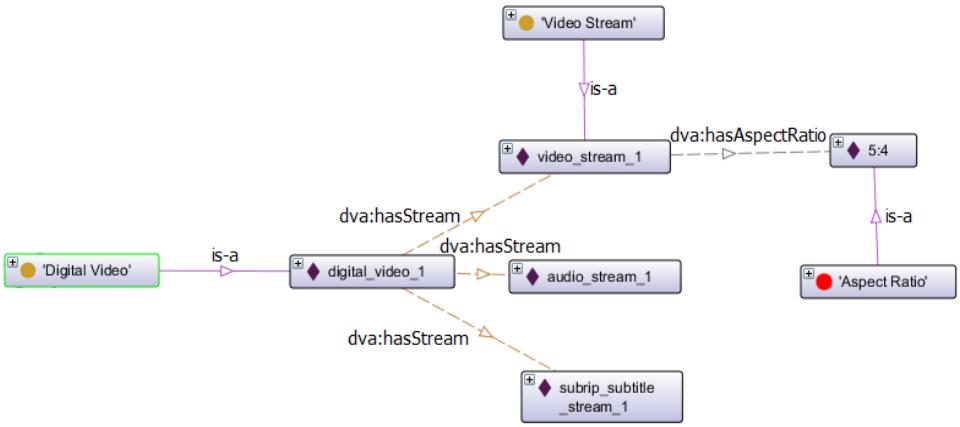

Fig. 5-9. Instances classified either manually (yellow circles), or automatically (red circle) through the inference of a reasoner that is based on relevant instances, relations and their values.

The applicability and efficiency of ontological inference methods, based on either class/property declarations (axioms) or SPIN rules, is highly dependent on the volume of information being stored in the examined ontology. The added knowledge can potentially lead to better classification results, since, in other words, the less the information stored within an ontology the less the axioms/rules that are met in reasoning process.

PROPheT can be the proper tool to be deployed for enriching instances in an ontology, so as to perform reasoning on the enriched instances. For example, as presented in sub-section "*Inference based on Domain and Range Restrictions*", if we have an instance of an unknown class that is connected with an instance of a `dva:VideoStream` class via property `dva:hasAspectRatio`, then the aforementioned instance will be classified to the `dva:AspectRatio` class (see Fig. 5-9).

[39] Description available at: http://mklab.iti.gr/project/prophet-ontology-populator, source code available at: https://github.com/MKLab-ITI/prophet
[40] http://wiki.dbpedia.org/
[41] http://www.europeana.eu/portal/en

Apart from enriching the content of an ontology, our tool can also instantiate information about the *use-context* of instances in the ontology. By the term use-context we consider any relevant information regarding the "context of use" of entities (digital objects) in their environment [Kontopoulos et al., 2016]. As already stated in [PERICLES D4.3, 2016], the representation of use-context capitalises on the LRM notions defined as dependency descriptors. Through a relevant parameter selection in PROPheT, the tool gives the ability to create links between populated instances and instances from external sources. An indicative description of the triples that are automatically created in a sample population process of a single instance, are given below in RDF Turtle[42] form:

```
@prefix     my_source: <http://PROPheT_sample_ontology#> .
@prefix     external_source: <http://external_sample_ontology#> .
@prefix     rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix     lrm: <http://xrce.xerox.com/LRM#> .

my_source:dependency_xyz rdf:type   lrm:Dependency .
my_source:dependency_xyz lrm:from   my_source:populated_instance_klm .
my_source:dependency_xyz lrm:to     external_source:instance_abc .
```

where `dependency_xyz` is the new instance of dependency created, that links the newly `populated_instance_klm` with data derived from the `instance_abc` of the external LOD source. As stated in Section "Inference Based on Class Axioms", if more details are stored in the ontology regarding the `lrm:from` part of each dependency, then further classification of the instance of dependency into `HardwareDependency`, `SoftwareDependency` or `DataDependency` class, will be feasible.

5.2.5. Contextualised Reasoning on Semantic Drifts

This section presents a method for performing contextualised reasoning utilizing semantic drift measures as a means of contextual information. **Semantic drift** is an area of active research, related to ontology evolution. It investigates the phenomenon of change in the meaning of concepts, possibly even overlapping and overtaking the meaning of other concepts within knowledge representation models, usually over time. This phenomenon can have drastic consequences on the use of knowledge representation models in applications and, therefore, metrics for its assessment and characteristics are valuable to knowledge engineering experts.

The previous deliverable, D4.4, presented a set of developed metrics for measuring drift. Additions and formalization to these definitions are presented here. In this deliverable we extend our previous work by utilizing the developed metrics and methods for contextualised reasoning. For this purpose we initially develop a model to store and manage drift metrics within an ontology, which is presented in the "*Modelling Drift Knowledge: The Drift Ontology*" section. Well-defined knowledge about concept drift can enable further reasoning and insights to discover volatile entities within a model and grasp its overall consistency. Such methods are presented in section "*Reasoning on Drift Knowledge*". Finally, the tools and GUIs developed to further harness the proposed methods in a user-friendly manner are presented in section "*Tools and Applications for Semantic Drift*". The tools are platform-independent and directly applicable to any model and domain, promoting the dissemination of the project's work and outcomes beyond Digital Preservation.

[42] https://www.w3.org/TR/turtle/

Background

A framework for measuring semantic drift in two or more ontology versions was presented in D4.4, along with a disambiguation of terms in the field, such as semantic change, concept drift, semantic decay and shift. The metrics presented are mainly focused around three aspects and comply with two approaches, namely identity- and morphing-based. In detail, the aspects are:

- *Label*, which refers to the description of a concept, via its name or title;
- *Intension*, which refers to the characteristics implied by it, via its properties;
- *Extension*, which refers to the set of things it extends to, via its number of instances.

A formal definition of the terms, building upon D4.4, is as follows:

$$label_t(C) = \{l, \mid \forall \langle C, rdfs{:}label, l \rangle \in T\}$$

$$int_t(C) = \{i \mid \ i = \langle C, p, x \rangle \vee \ i = \langle x, p, C \rangle, p = rdfs{:}domain \vee p = rdfs{:}range, \forall i \in T\}$$

$$ext_t(C) = \ \{x \mid \forall \langle x, rdf{:}type, C \rangle \ \in T\}$$

where *T* is the set of all triples in the ontology version *t*. In other words:

- The label aspect is given by the `rdfs:label` of a concept.
- The intension aspect is a set comprised of the union of all RDF triples with *C* in the subject or object position of OWL Object Properties or OWL Datatype Properties.
- The extension aspect is defined as the set of all instances of `rdf:type` *C*.

The two approaches for measuring drift refer to the assumption that the chain of corresponding concept identities across versions is either known or unknown. Much philosophical debate is involved in how and by which properties can identify a concept across time and how this can be formalized [Guarino & Welty, 2000]. Some approaches utilize the notions of perdurance and endurance, as defined in [Gangemi et al., 2002], to seek identity, such as by looking at rigid, properties that have to be persistent for all instances of a concept [Meroño-Peñuela & Hoekstra, 2014]. Finally, we consider two approaches as introduced in [Wang et al., 2011]:

- **Identity-based approach** (i.e. known concept identity): Assessing the extent of shift or stability of a concept's meaning is performed under the assumption that its identity is known across ontologies. For instance, considering an ontology A, and its evolution, ontology B, each concept of A is known to correspond to a single, known concept of B.
- **Morphing-based approach** (i.e. unknown concept identity): Each concept is pertaining to just a single moment in time (ontology), while its identity is unknown across versions (ontologies), as it constantly evolves/morphs into new, even highly similar, concepts. Therefore, its change has to be measured in comparison to every concept of an evolved ontology.

While the implementation of metrics for the different aspects remains generally applicable, the current proposed method adopts and follows the **morphing-based approach**. Despite several methods have been proposed to seek identity correspondence across versions [Meroño-Peñuela & Hoekstra, 2014], they still can be domain or model dependent, mandating for ad-hoc expert knowledge in the form of annotations, user input or using explicit identities. For this novel method to remain as generally-applicable as possible, without prior processing and user input, we follow the morphing-based approach, assuming each concept morphs into a new highly similar one in each version. Drift is, hence, measured as the dissimilarity of two maximally similar concepts in two versions [Wang et al., 2011].

Modeling Drift Knowledge: The Drift Ontology

The *Drift Ontology* is a model of concept drift measures between two concepts or two versions of a concept in two models (e.g. changes in time). Its main concept is `ConceptDrift`, which represents a generic metric of semantic drift between two concepts. The class has one Datatype Property of name

`value` and numeric range to represent the measure of drift. Most often, this value measures similarity and ranges between zero and one, although the model does not enforce such constraint. The objects, for which the change is measured, are stored using the Object Properties `from` and `to` (i.e. class `ConceptDrift` is an extension of LRM's Dependency class). Notably, the means/algorithms used to actually measure change are transparent to the model. Therefore, it does not register whether directionality matters, but it is able to support it if it does. In case directionality matters, the Drift Ontology model can handle it by referring to the former concept (e.g. an earlier version) with `from` and to the latter concept (e.g. a subsequent version) with the property `to`. Meanwhile, if directionality does not matter, the same properties `from` and `to` will be used, but in any order, which is up to the author/expert.

Three kinds of concept drift are defined in the Drift Ontology, according to [Wang et al., 2011]: *Label*, *Intension* and *Extension*. This is represented with respective subclasses of `ConceptDrift`: `LabelConceptDrift`, `IntensionConceptDrift` and `ExtensionConceptDrift`. Naturally, the classes inherit the `value` property, to store the metrics, and the `from` and `to` properties to store the subjects.

In practice, concept drift concepts are instantiated for measuring drift between two instances of existing concepts. Since the subjects we wish to refer to will generally be concepts themselves, the use of OWL's **punning**[43] is recommended, i.e. referring to a class as an instance itself (done by creating instances with their class type URI). The Drift Ontology model and an instance of its usage are displayed below, measuring the extensional, intensional and label concept drift between the concept of `'Mixed Media Artwork'` in the tate_2011 model and the `'Software Based Artwork'` in the tate_2012 model, as 0.17, 0.5 and 0.8 respectively.

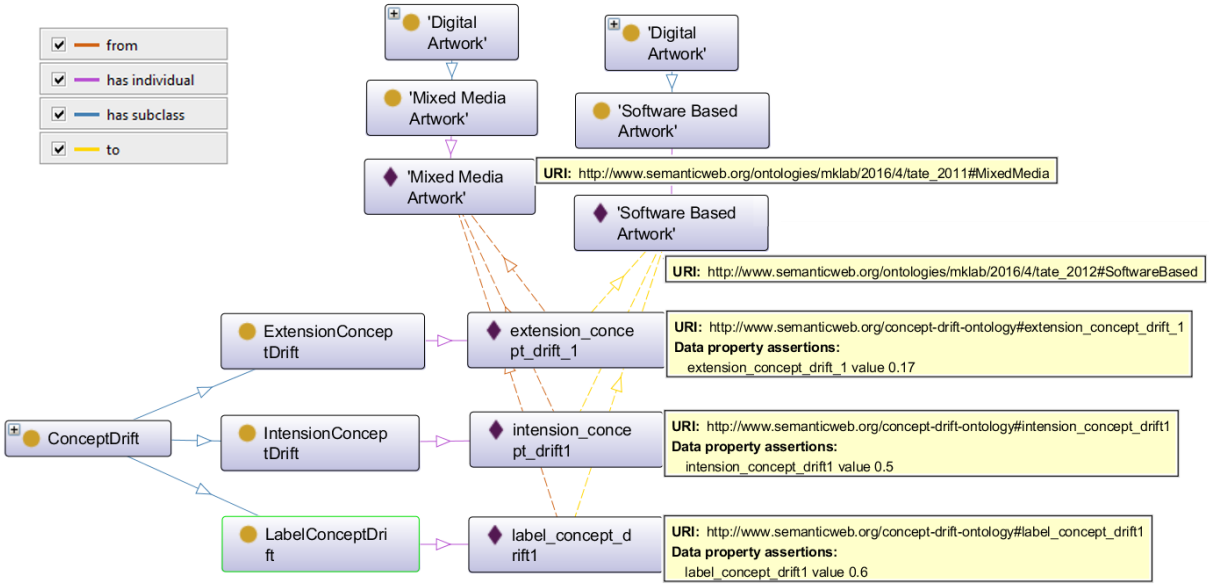

Fig. 5-10. The main concept drift ontology classes and example instances for digital artwork.

Reasoning on Drift Knowledge

After establishing a representation for the drift measurements within the Drift Ontology, reasoning capabilities can be utilized for various purposes and provide even further insight for knowledge models. The drift metrics essentially reveal properties of **concept stability over time**. With the aid of reasoning, such metrics can be interpreted as an inherent concept characteristic at any given

[43] https://www.w3.org/TR/owl2-new-features/#F12:_Punning

instance, i.e. any of these versions. Reasoning can, therefore, serve as a tool for ontology design and decision making. One goal to explore these capabilities is to identify **volatile entities**, i.e. concepts that are highly unstable and do not play a persistent role in the ontology.

The proposed method to pinpoint volatile entities involves the following steps:

1. For each version of the ontology and for each concept within, calculate the three drift aspect metrics, *label, extension* and *intension drift*.
2. Generate the corresponding entities.
3. Extract the overall drift metric, *whole drift*, for each concept.
4. Apply thresholds to each extracted *whole drift* to decide on a set of *volatile concepts* as a subset of all the concepts within the ontology.

Consequently the method will properly visualize the set of volatile concepts to enable evaluation. This will be done initially in terms of a table of volatile versus persistent concepts. Also, the hierarchy may be viewed with faded volatile entities so as to help with further model design and decision making. This feature can be easily linked up with the drift logs of the Somoclu analytical framework (see D4.4) and is included in our plans for future research.

Reasoning on drift metrics can also serve as means for consistency checking. Measuring and visualizing drift can further assure experts for the consistency of the ontology and help pinpoint conflicting concepts. For this purpose, we plan to provide reasoning combined with visualizing morphing chains of concept drift across time. Such a link could potentially convert statistical/distributional semantics-based drift monitoring to ontological/logical semantics based reasoning, which is considered an important convergence.

Tools & Applications for Semantic Drift

A critical goal underlying our methods for reasoning on semantic drift measures is their direct applicability and dissemination in the Semantic Web community, targeting the lack of user-friendly universal tools. For this reason, we have developed a suite of tools that help bring our methods and outcomes to a variety of platforms, use case scenarios, and experts. All tools of the **SemaDrift suite** are domain independent i.e. applicable to any model.

The fundamental tool of the suite is the SemaDrift Library, which is essentially the core API implementing the calculation of metrics. The suite is complemented by two different front-ends to the APIs functionality: the SemaDrift Protégé plugin, intended mainly for knowledge engineers, and the SemaDrift Desktop, intended for general usage.

SemaDrift Library API

The SemaDrift Library API is intended for developers who wish to implement metrics in their applications. It is implemented in Java and utilizes OWL-API[44] to parse the ontologies. By providing utilities such as getting a tree-like structure of loaded ontologies, API clients are free to develop their applications without any expertise of the underlying tools or re-importing dependencies.

The core SemaDrift Library implements the following functionality:

- Load a chain of multiple ontology versions.
- Retrieve the ontologies in tree-like structures.
- Calculate drift metrics for all aspects.

In detail the metrics are:

- Overall average concept stability per aspect for the whole ontology.

[44] http://owlapi.sourceforge.net/

- Concept overall stability: the average stability of a concept across versions[45].
- Concept-per-concept stability: comparing each concept to all the rest, providing the metrics to construct a complete morphing chain.

SemaDrift Protégé Plugin

This plugin provides integration with the popular ontology creation software Protégé[46], providing a GUI in its environment to calculate drift. It leverages the Java SemaDrift Library to provide drift metrics for two consecutive versions: one open in Protégé and a second ontology of choice. An important characteristic of this plugin is that it enables knowledge engineers to work on their models in the popular Protégé software while occasionally comparing them to other editions via the plugin without leaving the environment.

The main GUI of the SemaDrift Protégé Plugin is shown in Fig. 5-11. In this example, we consider two consecutive ontology versions from Tate, as constructed in the framework of D4.4, tate2011.owl and tate2012.owl, modeling digital artwork information. In this scenario, we assume the former ontology to be loaded first on Protégé, possibly for modification by the knowledge engineer working on it. By loading the SemaDrift plugin, the right panel comes forward, while the main ontology is still visible on the left pane. The plugin's panel initially provides a file browser to search for a second ontology to compare two by pressing "Load". By pressing the "Calculate" button, the tables below are filled with the metric outcomes.

Evidently, the plugin provides a quick method to view overall drift compared to another model (average drift). It also allows going into further detail. The tables of concept-per-concept stability may be copied and used for constructing a visual morphing chain in any other software of choice.

SemaDrift Desktop

This standalone application in the Java Platform is intended for use in Desktop computers, enabling drift measurement between two consecutive ontology versions of choice. It provides a user-friendly GUI for leveraging the SemaDrift Library API. This tool complements the Protégé plugin for a variety of use case scenarios. The application is intended for a broader audience, such as non-experts who do not wish additional software installation, but merely a targeted solution for viewing drift. Leaving the Protégé environment, the tool allows for slightly more flexibility regarding visually appealing graphics and the addition of a second ontology tree visualization for browsing changes.

Fig. 5-12 shows the tool's main interface. From left to right, the tool enables loading two ontologies locally, using a file browser. Consequently, it generates a tree-like hierarchy to examine them in terms of classes and properties. In the event of changing the model in an external editor, the ontology can simply be reloaded. In the right panel, metrics are calculated on demand in a similar manner and form as in the Protege plugin. They include overall average drift, average concept drift by detecting identity and concept-per-concept drift tables to construct morphing chains.

[45] Notably, this metric is only possible if the identity of a concept is persistent and also known across versions, which contradicts to the morphing-based approach. It may also require knowledge of the model or input by a user which again contradict the general-purpose intention of the tools. However, we have devised a novel method to seek identity by regarding the most similar concept as the identical concept and use it to find average concept stability without the need for assumptions or user input.

[46] Protégé - http://protege.stanford.edu/

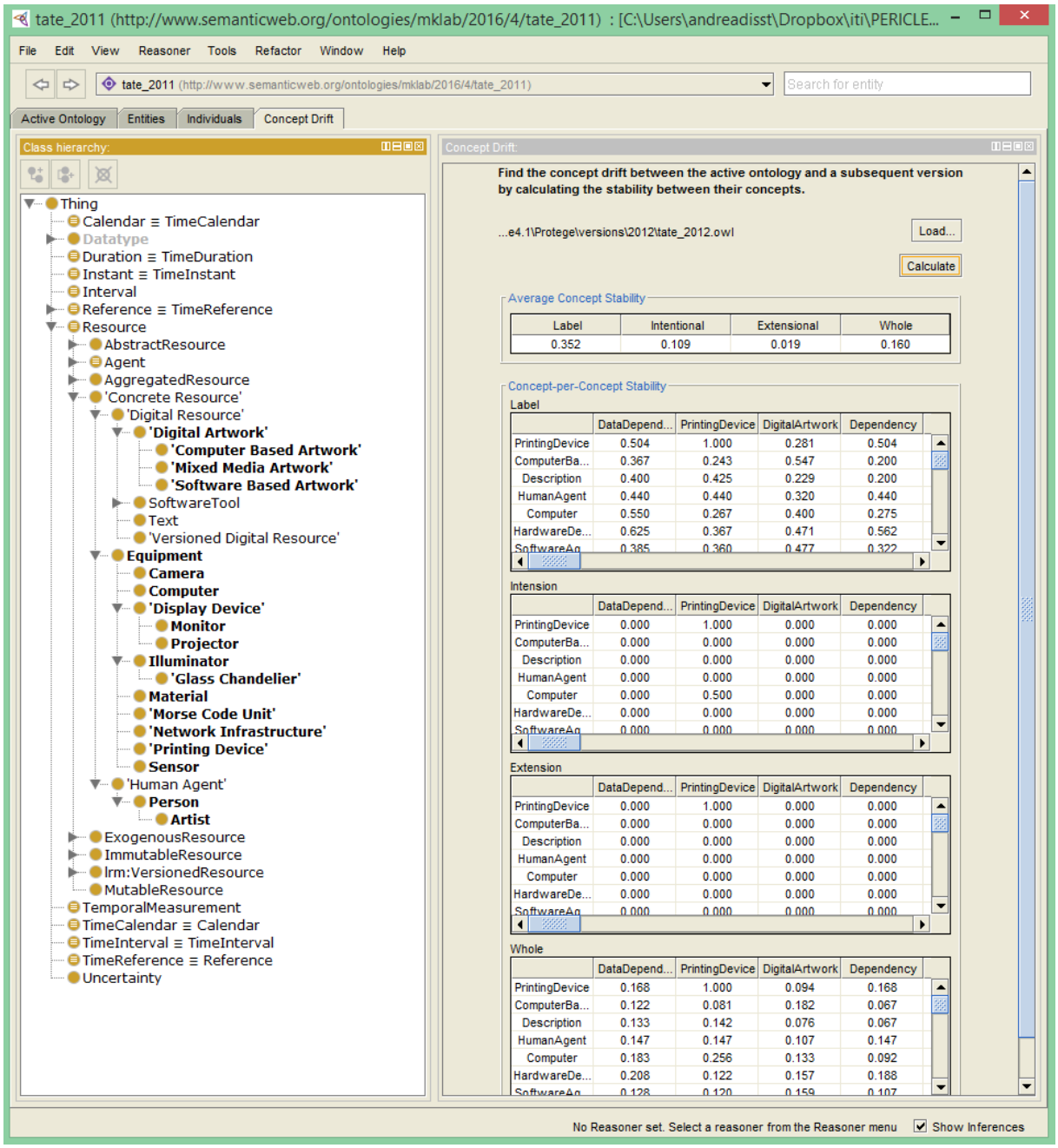

Fig. 5-11. SemaDrift Protégé Plugin showing drift metrics for two ontologies.

Future additions would include a dynamic pane which would support adding more than two ontologies, as already supported by the underlying SemaDrift Library to calculate the metrics. Also, ontologies could be loaded from the Web, searched for specific concepts and reloaded automatically when changed externally. The visualization can also be enhanced by adding a graphical morphing chain to results.

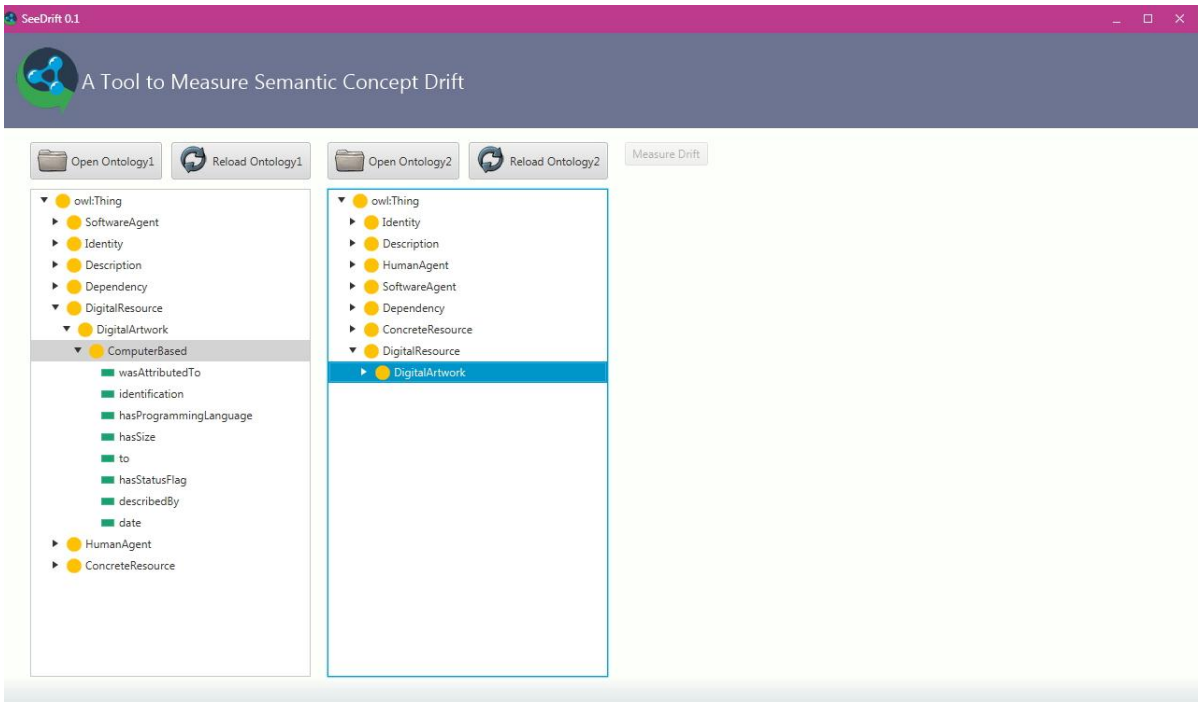

Fig. 5-12. SemaDrift Desktop software, showing two ontologies loaded.

5.3. Uncertainty Handling

Uncertainty is a parameter that unavoidably emerges in everyday reasoning and in real-world domains. The main sources of uncertainty include:

- **Incomplete knowledge**, like e.g. missing information and non-exhaustive modelling of domain knowledge;
- **Imprecise knowledge**, like eg. the time that an event happened can be known only approximately;
- **Unreliable knowledge**, like e.g. measurements coming from sensors that can be biased or defective.

In the case of ontology-based knowledge representation, like e.g. in PERICLES, the pertinent ontological commitments contain facts that either hold or do not hold in the domain of discourse. In order to represent uncertainty in the ontologies and the Semantic Web world, there have been many proposals for extensions of the underlying languages (RDF, OWL, DLs, rules) by uncertainty and vagueness, as indicated in the next subsection.

5.3.1. Approaches for Handling Inconsistent & Missing Knowledge

Probabilistic Ontologies

This direction of research refers to generalizing classical ontologies by probabilistic knowledge [Lukasiewicz, 2008]. The encoded probabilistic knowledge involves the representation of terminological probabilistic knowledge about concepts and roles (e.g. "*Birds fly with a probability of at least 0.95*"), and assertional probabilistic knowledge about instances of concepts and roles (e.g. "*Tweety is a bird with a probability of at least 0.9*"). Relevant applications of probabilistic ontologies include deployments in medicine, biology, defense and astronomy, while in the Semantic Web the main applications are in information retrieval, personalization, recommender systems and ontology matching [Ding & Peng, 2004; Giugno & Lukasiewicz, 2002].

Fuzzy DLs

Description logics model a domain of interest in terms of concepts and roles, which represent classes of individuals and binary relations between classes of individuals, respectively. A description logic knowledge base encodes in particular subset relationships between concepts, subset relationships between roles, the membership of individuals to concepts, and the membership of pairs of individuals to roles. In fuzzy description logics, these relationships and memberships then have a degree of truth in [0, 1], thus, whether an instance belongs to a concept is usually not a matter of "yes/no", but a matter of degree of membership.

The resulting fuzzy vagueness can be used for expressing vague concepts. In essence, fuzzy logic reflects the impression of human language and reasoning – examples of frequent fuzzy concepts are "young", "furniture", "most", "cloudy", and so on. A major difference between this approach and probabilities discussed above is the fact that uncertainty in fuzzy concepts usually does not get reduced with the coming of new information. Typically, in building a fuzzy system, the designer needs to provide all membership functions included in it, by considering how the concepts are used by average people. Most successful applications of fuzzy logic so far are in fuzzy control systems, where expert knowledge is coded into fuzzy rules [Jantzen, 2007; Tanaka & Wang, 2001].

Finally, fuzzy approaches face the following main challenges: (a) the degree of membership is often context dependent, and, (b) the general-purpose fuzzy rules are hard to get.

Non-monotonic Logics

A reasoning system is **monotonic** if the truthfulness of a conclusion does not change when new information is added to the system – the set of theorems can only monotonically grow when new axioms are added. In contrast, in **non-monotonic reasoning** systems, the set of conclusions may either grow or shrink when new information is obtained [Brewka, 1991; Horty, 2001].

Nonmonotonic logics are used to formalize plausible reasoning, such as the following inference step:

1. *Birds typically fly*.
2. *Tweety is a bird*.
3. *Tweety (presumably) flies*.

Such reasoning is characteristic of commonsense reasoning, where default rules are applied when case-specific information is not available.

The conclusion of nonmonotonic argument may turn out to be wrong. For example, if Tweety is a penguin, it is incorrect to conclude that Tweety flies. Nonmonotonic reasoning often requires jumping to a conclusion and subsequently retracting that conclusion as further information becomes available.

All systems of nonmonotonic reasoning are concerned with the issue of consistency. Inconsistency is resolved by removing the relevant conclusion(s) derived previously by default rules. Simply speaking, the truth value of propositions in a nonmonotonic logic can be classified into the following types:

1. *facts* that are definitely true, such as "Tweety is a bird"
2. *default rules* that are normally true, such as "Birds fly"
3. *tentative conclusions* that are presumably true, such as "Tweety flies"

When an inconsistency is recognized, only the truth value of the last type is changed.

Major problems in these approaches are (a) conflicts in defaults, such as in the "Nixon Diamond"[47], and, (b) computational expense: to maintain the consistency in a huge knowledge base is hard, if not impossible.

[47] https://en.wikipedia.org/wiki/Nixon_diamond

5.3.2. Rule-based Uncertainty Management

The uncertainty management framework in PERICLES is based on **defeasible logics** [Nute, 1994], a non-monotonic logics formalism that is extremely suitable for handling conflicts and uncertainty in information that is heterogeneous, diverse and possibly inconsistent. Defeasible logics can offer a flexible and human-intuitive formalism for efficiently handling such situations, since they feature a sophisticated conflict resolution mechanism implemented through a binary rule superiority relationship. Conflicts between two rules stem from complementary rule heads or heads with conflicting literals (i.e. pairs of mutually exclusive literals that cannot both be derived at the same time).

A defeasible theory *D* (i.e. a program written in defeasible logics) is a couple (*R*,>) where *R* is a finite set of rules and > a superiority relation on *R*. Each rule has a unique rule label. There are three kinds of rules: strict rules, defeasible rules, and defeaters:

- Strict rules, which are denoted by $A \longrightarrow p$, where *A* is a set of literals and *p* is a (positive or negative) literal, and are interpreted in the typical sense: whenever the premises are indisputable, then so is the conclusion. An example of a strict rule is "*Penguins are birds*". Written formally: r_1: *penguin*(*X*) $\longrightarrow$ *bird*(*X*). Inference from strict rules only is called definite inference. Strict rules are intended to define relationships that are definitional in nature and such an example is ontological knowledge.
- Defeasible rules are denoted by $A \Rightarrow p$, and can be defeated by contrary evidence. Two examples of such rules are r_2: *bird*(*X*) $\Rightarrow$ *flies*(*X*) (i.e. "*Birds typically fly*") and r_3: *penguin*(*X*) $\Rightarrow \neg$*flies*(*X*) (i.e. "*Penguins typically do not fly*").
- Defeaters are denoted as $A \rightsquigarrow p$ and are not used to actively support conclusions, but only to prevent some of them. An example of such a defeater is: r_4: *heavy*(*X*) $\rightsquigarrow \neg$*flies*(*X*), which reads as: "*Heavy birds may not fly*".

A superiority relation on R is an acyclic relation > on R (that is, the transitive closure of > is irreflexive). When $r_1 > r_2$, then r_1 is called superior to r_2, and r_2 inferior to r_1. This expresses that r_1 may override r_2. For example, given the defeasible rules r_2 and r_3 above, no conclusive decision can be made about whether a penguin flies, because rules r_2 and r_3 contradict each other. But if we introduce a superiority relation > with $r_3 > r_2$, then we can indeed conclude that a penguin does not fly.

The defeasible reasoning layer in PERICLES is placed "over" the domain ontologies' level, performing reasoning on the knowledge stored in the ontologies. The implementations of the defeasible theories presented here are based on SPINdle, a popular defeasible logic rule engine [Lam & Governatori, 2009], while for authoring the rules we used SPINdle's online demo editor[48]. The interested reader can also refer to S^2DRREd (Syntactic-Semantic Defeasible Reasoning Rule Editor) [Kontopoulos et al., 2012], a more flexible rule authoring tool.

5.3.3. An Uncertainty Management Example: Impacted & Unimpacted DOs

This example is adopted from the way MICE (the Model Impact Change Explorer component) determines resources that are impacted or unimpacted by a change taking place in a Digital Ecosystem entity. MICE allows the user to register a change in the ecosystem and then view the impacted and unimpacted resources through a visualized dependency graph, showing the impact that the change might have to other resources in the ecosystem.

Consider the following scenario adapted from the DVA Ontology:

[48] available at: http://spin.nicta.org.au/spindle/demo.html

- Digital Video 1 is connected to an AVI container, through property hasContainer.
- The AVI container has a software dependency (software_dependency_1) from certain media players: Windows Media Player 12, QuickTime Player and VLC Media Player.
- Windows Media Player 12 has a compatibility dependency (software_dependency_3) from certain operating systems: Windows 7, Windows 8 and Windows 10.
- Windows 7 has a compatibility dependency (hardware_dependency_2) from certain resources: 1GB RAM and 1GB Processor (these are the minimum requirements for installing/running Windows 7).

The above dependencies form a chain (i.e. "chain of dependencies"), which is illustrated in Fig. 5-13 below:

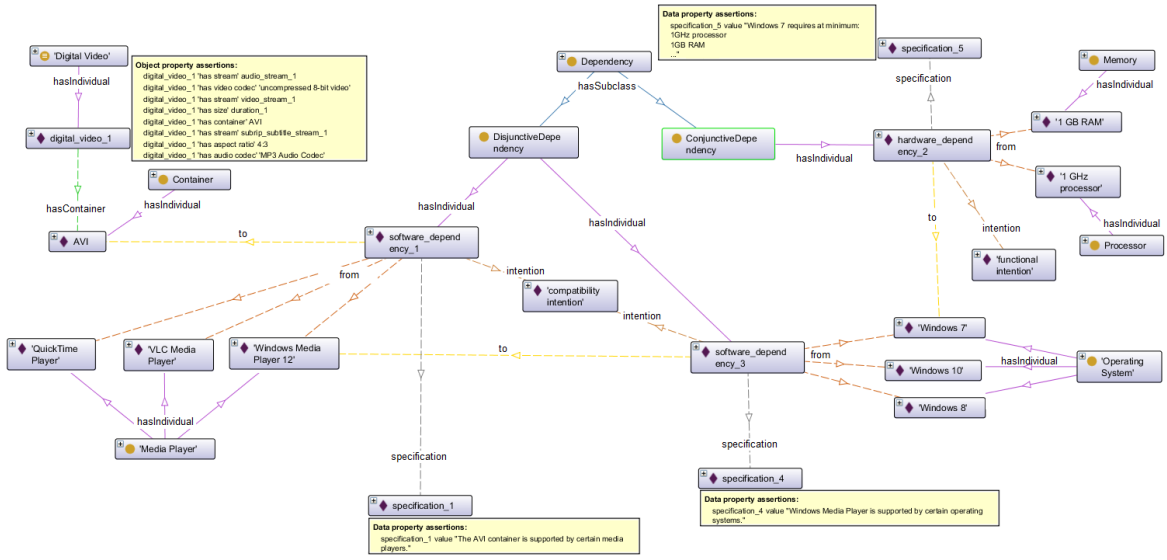

Fig. 5-13. "Chain of dependencies" scenario from the DVA Ontology.

As shown in the figure, there are two types of LRM dependencies [PERICLES D3.3, 2015]:

- **Conjunctive Dependencies**: ALL "from" resources should be consistent in order for the "to" resource to be consistent.
- **Disjunctive Dependencies**: ANY of the "from" resources should be consistent in order for the "to" resource to be consistent.

The above principles for determining when a resource is impacted or not can be expressed with the following defeasible logic rule base:

r_{11}: conjunctive(D), from(D,X), to(D,Y), impacted(X) ⇒ impacted(Y)
r_{21}: disjunctive(D), from(D,X), to(D,Y), impacted(X) ⇒ impacted(Y)
r_{22}: disjunctive(D), from(D,X1), from(D,X2), to(D,Y), impacted(X1), ¬impacted(X2), X1≠X2 ~> ¬impacted(Y)

Rule r_{11} states that whenever a resource in the "from" part of a conjunctive dependency is impacted, the impact is *most probably* transferred to the "to" part as well. This is a defeasible rule, meaning that it can be defeated in case contrary evidence is presented. Quite similarly, rule r_{21} states that whenever all resources in the "from" part of a disjunctive dependency are impacted, the impact is probably transferred to the "to" part as well. However, in this case rule r_{22} is the exception to r_{21}, stating that if at least one resource exists in the "from" part of a disjunctive dependency that has not been impacted, then the resource in the "to" part is not impacted either. Note that rule r_{22} is a defeater, meaning that it can only defeat r_{21}'s defeasible conclusion. In practice it implements the exception to rule r_{21} (exceptions are characteristic of the flexibility provided by defeasible logics).

Having the above dependency chain and starting with an impacted “1GHz Processor” (e.g. the processor is removed or changed), one can experiment with various cases, like the ones below:

- Hardware Dependency 2 is a conjunctive Dependency, while Software Dependency 3 and Software Dependency 1 are disjunctive dependencies. In this case, the “Windows 7” resource is impacted, besides the originally impacted “1GHz Processor”.
- Hardware Dependency 2, Software Dependency 3 and Software Dependency 1 are Disjunctive dependencies. In this case, only the “1GHz Processor” resource is impacted.
- Hardware Dependency 2 and Software Dependency 3 are Conjunctive dependencies, while Software Dependency 1 is a Disjunctive dependency. In this case, “Windows 7” and “Windows Media Player 12” are also impacted.

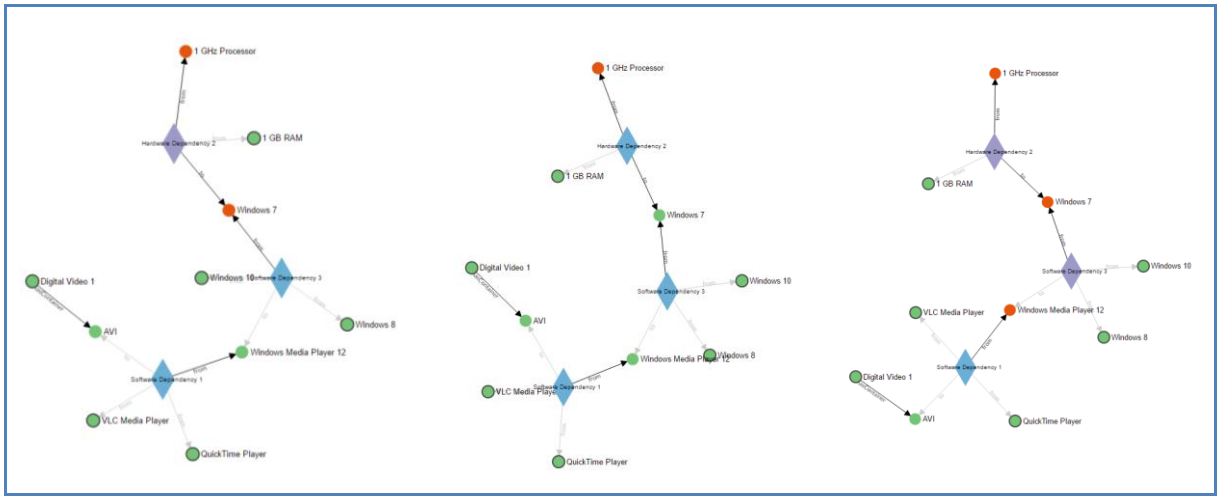

Fig. 5-14. Dependency graphs for the above cases.

Below is the defeasible logic rule base for the above cases in SPINdle syntax. Note that SPINdle currently lacks a theory grounding mechanism (i.e. no variables can be used in the predicates), which would be greatly beneficial for the purposes of this work. However, the engine’s advantages (fast, reliable, highly integrated and easily-deployable) constitute the incentive for preferring the specific reasoner. The lack of grounding mechanism is solved in our approach by replacing atoms containing arguments and variables with appropriately composed sets of synthetic predicates, as seen below:

```
# software dependency 1
>> sw_dep_1
>> from_swdep1_WMP
>> from_swdep1_QTP
>> from_swdep1_VLC
>> to_swdep1_AVI

# hardware dependency 2
>> hw_dep_2
>> from_hwdep2_1GHzProcessor
>> from_hwdep2_1GBRAM
>> to_hwdep2_Win7

# software dependency 3
>> sw_dep_3
>> from_swdep3_Win7
>> from_swdep3_Win8
>> from_swdep3_Win10
>> to_swdep3_WMP

d11: impacted_WMP, from_swdep1_WMP, to_swdep1_AVI, disj_dep_1
      ~> -impacted_AVI
```

```
d12: impacted_QTP, from_swdep1_QTP, to_swdep1_AVI, disj_dep_1
      ~> -impacted_AVI
d13: impacted_VLC, from_swdep1_VLC, to_swdep1_AVI, disj_dep_1
      ~> -impacted_AVI
d14: impacted_WMP, impacted_QTP, impacted_VLC, from_swdep1_WMP,
from_swdep1_QTP, from_swdep1_VLC, to_swdep1_AVI, disj_dep_1
      => impacted_AVI
d15: impacted_WMP, from_swdep1_WMP, to_swdep1_AVI, conj_dep_1
      => impacted_AVI
d16: impacted_QTP, from_swdep1_QTP, to_swdep1_AVI, conj_dep_1
      => impacted_AVI
d17: impacted_VLC, from_swdep1_VLC, to_swdep1_AVI, conj_dep_1
      => impacted_AVI

d14 > d11
d14 > d12
d14 > d13

d21: impacted_1GHzProcessor, from_hwdep2_1GHzProcessor, to_hwdep2_Win7,
disj_dep_2
      ~> -impacted_Win7
d22: impacted_1GBRAM, from_hwdep2_1GBRAM, to_hwdep2_Win7, disj_dep_2
      ~> -impacted_Win7
d23: impacted_1GHzProcessor, impacted_1GBRAM, from_hwdep2_1GHzProcessor,
from_hwdep2_1GBRAM, to_hwdep2_Win7, to_hwdep2_Win7, disj_dep_2
      => impacted_Win7
d24: impacted_1GHzProcessor, from_hwdep2_1GHzProcessor, to_hwdep2_Win7,
conj_dep_2
      => impacted_Win7
d25: impacted_1GBRAM, from_hwdep2_1GBRAM, to_hwdep2_Win7, conj_dep_2
      => impacted_Win7

d23 > d21
d23 > d22

d31: impacted_Win7, from_swdep3_Win7, to_swdep3_WMP, disj_dep_3
      ~> -impacted_WMP
d32: impacted_Win8, from_swdep3_Win8, to_swdep3_WMP, disj_dep_3
      ~> -impacted_WMP
d33: impacted_Win10, from_swdep3_Win10, to_swdep3_WMP, disj_dep_3
      ~> -impacted_WMP
d34: impacted_Win7, impacted_Win8, impacted_Win10, from_swdep3_Win7,
from_swdep3_Win8, from_swdep3_Win10, to_swdep3_WMP, disj_dep_3
      => impacted_WMP
d35: impacted_Win7, from_swdep3_Win7, to_swdep3_WMP, conj_dep_3
      => impacted_WMP
d36: impacted_Win8, from_swdep3_Win8, to_swdep3_WMP, conj_dep_3
      => impacted_WMP
d37: impacted_Win10, from_swdep3_Win10, to_swdep3_WMP, conj_dep_3
      => impacted_WMP

d34 > d31
d34 > d32
d34 > d33

>> impacted_1GHzProcessor
>> impacted_1GBRAM

# case 1
# s1: sw_dep_1 -> disj_dep_1
# s2: hw_dep_2 -> conj_dep_2
# s3: sw_dep_3 -> disj_dep_3
```

```
# case 2
# s1: sw_dep_1 -> disj_dep_1
# s2: hw_dep_2 -> disj_dep_2
# s3: sw_dep_3 -> disj_dep_3

# case 3
# s1: sw_dep_1 -> disj_dep_1
# s2: hw_dep_2 -> conj_dep_2
# s3: sw_dep_3 -> conj_dep_3
```

5.3.4. Other Uncertainty Management Applications in PERICLES

Some other cases in PERICLES where uncertainty management could be deployed include the various types of classification tasks and consistency checks described earlier in this chapter, like e.g.:

- **Classifying an instance** as a DVA, SBA or BDA under uncertainty.
- **Classifying a dependency** into one of the specialized types (e.g. hardware dependency, etc.).
- **Consistency checking** - classifying resources as 'warning' or 'error' items.

In all the above cases, one could extend the deterministic definitions in the ontologies with appropriate sets of defeasible logic rules that would handle the emergence of contradictory or inconsistent information. A theory grounding mechanism for SPINdle would be greatly beneficial in all these cases, in order to avoid creating cumbersome rule bases like the one presented in the previous subsection. There has been an initial attempt towards this aim [Rohaninezhad et al., 2015], but a more sophisticated solution is still in the works[49].

5.4. Chapter Summary

In this chapter we described a framework for deriving useful interpretations and for inferring logical consequences from a set of asserted facts or axioms stored in ontological models. The derived facts are not explicitly stated in the ontology, nevertheless the knowledge, the semantics and the contextualized content of the domain of interest is stored in it in a formal way. The described PERICLES semantic interpretation framework is based on logical semantics, on established reasoning engines and ontological inference techniques. We demonstrate how to infer implicit knowledge based on ontology declarations (class axioms, properties' characteristics, and domain/range restrictions). Advanced reasoning techniques, through the implementation of SPIN rules (SPARQL Inferencing Notation), are presented and additionally detect inconsistencies while examining a specific state of a digital ecosystem. We have also proposed the implementation of SPIN rules in a way to monitor policies and handle changes, as expressed in specific LRM instantiations in the examined model. Furthermore, we described the method and the developed tools for contextualised reasoning by utilizing semantic drift measures as a means of contextual information. Contextualized reasoning on semantic drifts offers the capability of determining the "volatile" and conflicting concepts in an ontology model. Finally, in this chapter we outlined our proposed scheme for uncertainty management in contextualized content representations, which is based on non-monotonic and defeasible logics.

[49] We contacted the authors of [Rohaninezhad et al., 2015], who said that they are developing a RESTful version for the updated SPINdle rule engine featuring theory grounding.

6. Conclusions and Next Steps

6.1. Conclusions

This deliverable reported on the work conducted in T4.5, focusing on contextualised content interpretation and presenting our respective proposed approaches for gaining insightful contextualised views on content semantics. Our work is based on the context-aware semantic models developed within the project (WP2, WP3) and is aimed at deriving higher level contextualised content interpretations that are closer to human perception. The deliverable investigated our two diverse but complementary approaches – quantum-like analysis and semantic reasoning-based analysis – and presented the following outputs per topic:

- **Contextualised Content Interpretation**: The deliverable started with an introductory overview of our proposed schemes and motivation for representing contextualized content semantics, under the scope of two core approaches presented in the next chapters. A theoretical framework is presented using physics as a metaphor to develop different models of evolving semantic content. Based on this framework, we are making a big step towards modelling the dynamics of semantic drifts on language based "forces".
- **Quantum-Like Analysis for Contextual Content Interpretation**: The deliverable discussed the quantum-like nature of semantic content related user behaviour and presented the implementation of a quantum-like model for semantic content classification. The implementation investigated the move from context-dependence of semantic content to contextuality (non-commutativity) and entanglement, integrating two PERICLES tools, Somoclu for drift detection and Ncpol2spda for entanglement detection. The result is deriving a generalized "energetic" hypothesis underlying contextualized semantic content behaviour over time.
- **Semantic Reasoning for Contextual Content Interpretation**: The document also reported on our second line of research on contextualized content interpretation based on logical semantics and involving the use of semantic reasoning techniques. The proposed PERICLES semantic interpretation framework capitalizes on our adopted representations for contextualized content semantics (see D4.4) and integrates: (a) an ontological inference scheme based on Description Logics; (b) a SPIN-based rule reasoning layer; (c) an uncertainty management framework based on defeasible logics, a member of the non-monotonic logics family. Additional components included are: a novel scheme for contextualized reasoning on semantic drift, accompanied by respective tool implementations (the SemaDrift API, desktop application and Protégé plugin), and, an implementation of SPIN rules for policy and ecosystem change management, based on the LRM representation of preconditions and impacts.
- All developed models and software tools for the investigations within D4.5 are publicly available along with the respective results and datasets.

6.2. Next Steps

Although our RTD activities in WP4 are concluded with this deliverable, there are a series of possible next steps beyond PERICLES in different but related directions, with their outcome contributing to LTDP by new challenges and technologies as follows:

- Integration of vector field based word semantics with the LRM and QT-based computational linguistics, prominently e.g. [Blacoe et al., 2013; Coecke et al., 2010; Cohen et al., 2010]. This is an important track as on the one hand, it extends the field model from word to phrase and sentence semantics in order to couple ontologies with statistics. On the other hand there is another

speedup available by coupling the outcome with QML [Wittek, 2014] and quantum computing [Zeng & Coecke, 2016];

- Based on the above, integration of the semantic reasoning component with the vector field approach to evolving semantics for advanced knowledge representation. An already available link for robust reasoning by semantic vectors is [Widdows & Cohen, 2015];
- To represent logical statements in vector space or in a vector field constitutes a future research track for the second part of this deliverable. More importantly, it paves the way for the **RDF-based** (Resource Description Framework) **indexing of DOs**, and thereby for using the Linked Resource Model (LRM) in a new environment with statistical foundations;
- Work on an EM-like "dipole" representation of semantic content to add a second "force" besides gravity to the representation toolbox. This will realize the mapping of evolving semantic content in scalable distributed/virtual collections according to two complementary paradigms, and as a result will consider the Semantic Web as a global knowledge repository combined with the ability to reason over its holdings, practically a cognitive layer wrapping the planet. In this next level of content morphologies, we expect future research to identify relatively stable semantic constellations similar to galaxies (Fig. 6-1).

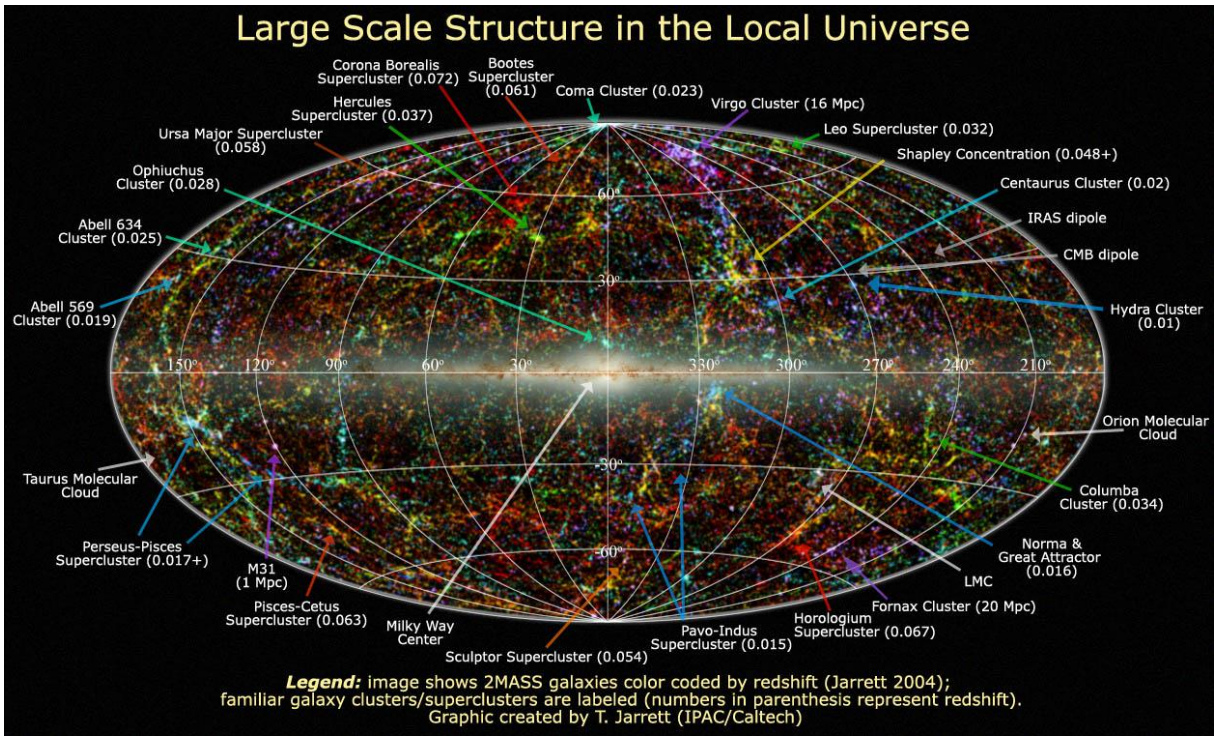

Fig. 6-1. Constellations with names in the local Universe as a prototype for topically related content galaxies [Jarrett, 2004].

7. References

[Acciarri et al., 2005] Acciari, A., Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Palmieri, M., Palmieri, M., and Rosati, R. (2005). QUONTO: querying ontologies. In AAAI, Vol. 5, pp. 1670-1671.

[Adams et al., 2012] Adams, G.K., Watson, K.K., Pearson, J., and Platt, M.L. Neuroethology of Decision-Making. Current Opinion in Neurobiology. 2012; 22(6):982-989.

[Adar et al., 2008] Adar, E., Dontcheva, M., Fogarty, J., and Weld, D. 2008. Zoetrope: interacting with the ephemeral web. In Proc. of the 21st ACM Symp. User Interf. Soft. and Tech. USA.

[Aerts & Sozzo, 2016] Aerts, D., and Sozzo, S. From Ambiguity Aversion to a Generalized Expected Utility. Modeling Preferences in a Quantum Probabilistic Framework. Journal of Mathematical Psychology.

[Aerts et al., 2006] Aerts, D., Czachor, M., and D'Hooghe, B. 2006. Towards a quantum evolutionary scheme: violating Bell's inequalities in language. In N. Gontier, J. P. Van Bendegem and D. Aerts (Eds.) Evolutionary Epistemology, Language and Culture. Amsterdam: John Benjamins. 453-478.

[Aghajanyan, 2015] Aghajanyan, A. 2015. Introduction to Gravitational Clustering. Pattern Analysis and Machine Intelligence X(Y), February 2015. At https://arxiv.org/pdf/1509.01659.pdf

[Alam et al., 2016] Alam, S. + 71 authors. 2016. The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological analysis of the DR12 galaxy sample. At http://arxiv.org/pdf/1607.03155v1.pdf

[Alonso et al., 2011] Alonso, O., Strötgen, J., Baeza-Yates, R., and Gertz, M. 2011. Temporal Information Retrieval: Challenges and Opportunities. In Proceedings of the International Temporal Web Analytics Workshop TWAW 2011, Hyderabad, India. 1-8.

[Amari & Wu, 1999] Amari, S., and Wu, S. 1999. Improving support vector machine classifiers by modifying kernel functions. Neural Networks, 12, 783-789.

[Asano et al., 2012] Asano, M., Basieva, I., Khrennikov, A., Ohya, M., Tanaka, Y., and Yamato, I. (2012). A quantum-like model of Escherichia coli's metabolism based on adaptive dynamics. In Proceedings of QI-12, 6th International Quantum Interaction Symposium, pages 60-67.

[Ashtiani & Azgomi, 2014] Ashtiani, M., and Azgomi, M.A. Contextuality, Incompatibility and Biased Inference in a Quantum-Like Formulation of Computational Trust. Advances in Complex Systems. 2014; 17(05):1450020.

[Ashtiani & Azgomi, 2015] Ashtiani, M., and Azgomi, M.A. A Survey of Quantum-like Approaches to Decision Making and Cognition. Mathematical Social Sciences. 2015; 75:49-80.

[Baader, 2003] Baader, F. (2003). The description logic handbook: Theory, implementation and applications. Cambridge University Press.

[Baker, 2008] Baker, A. 2008. Computational approaches to the study of language change. Language and Linguistics Compass 2, 3, 289-307.

[Bar-Ilan, 2008] Bar-Ilan, J. (2008). Informetrics at the beginning of the 21st century-a review. Journal of Informetrics, 2(1):1-52.

[Bateson, 1972] Bateson, G. (1972). Steps to an Ecology of Mind. New York: Ballantine, pp. Xxv-xxvi.

[Bechhofer, 2009] Bechhofer, S. (2009). OWL: Web ontology language. In Encyclopedia of Database Systems (pp. 2008-2009). Springer US.

[Beeferman et al., 1997] Beeferman, D., Berger, A., and Lafferty, J. 1997. A model of lexical attraction and repulsion. In Proceedings of ACL-97, Madrid, Spain. 373-380.

[Bell, 1964] Bell, J. (1964). On the Einstein-Podolsky-Rosen paradox. Physics, 1(3):195-200.

[Berners-Lee, 1998] Berners-Lee, T. (1998). Semantic web road map.

[Blacoe & Lapata, 2012] Blacoe, W., and Lapata, M. 2012. A Comparison of Vector-based Representations for Semantic Composition. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP 2012)

[Blacoe et al., 2013] Blacoe, W., Khasefi, E., and Lapata, M. 2013. A Quantum-Theoretic Approach to Distributional Semantics. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2013).

[Blacoe, 2015a] Blacoe, W. 2015. Semantic Composition Inspired by Quantum Measurement. In Proceedings of QI-14. pp 41-53.

[Blacoe, 2015b] Blacoe, W. 2015. A Tensor-Based DisCo Model for Sentence-Level Tasks. In Proceedings of the 1st Workshop on Advances in Distributional Semantics (ADS 2015)

[Bohm & Hiley, 1993] Bohm, D., and Hiley, B. (1993). The Undivided Universe: An Ontological Interpretation of Quantum Mechanics. Routledge and Kegan Paul, London.

[Bohm & Hiley, 1993] Bohm, D., and Hiley, B. (1993). The Undivided Universe: An Ontological Interpretation of Quantum Mechanics. Routledge and Kegan Paul, London.

[Borgman & Furner, 2005] Borgman, C. L. and Furner, J. (2005). Scholarly communication and bibliometrics. Annual Review of Information Science and Technology, 36(1):2-72.

[Börner et al., 2006] Börner, K., Penumarthy, S., Meiss, M., and Ke, W. (2006). Mapping the diffusion of scholarly knowledge among major US research institutions. Scientometrics, 68(3):415-426.

[Bornholt et al., 2016] Bornholt, J., Lopez, R., Carmean, D.M., Ceze, L., Seelig, G., and Strauss, K. (2016). A DNA-Based Archival Storage System. In Proceedings of ASPLOS '16, April 2–6, 2016, Atlanta, GA, USA.

[Brewka, 1991] G. Brewka (1991). Nonmonotonic Reasoning: Logical Foundations of Commonsense. Cambridge University Press.

[Brumby & Zhuang, 2015] Brumby DP, and Zhuang S. Visual Grouping in Menu Interfaces. In: Proceedings of CHI-15, 33rd Conference on Human Factors in Computing Systems. vol. 33; 2015. p. 4203-4206.

[Bruza & Woods, 2008] Bruza, P. and Woods, J. (2008). Quantum collapse in semantic space: interpreting natural language argumentation. In Proceedings of QI-08, 2nd International Symposium on Quantum Interaction.

[Bruza et al., 2009] Bruza, P.D., Widdows, D., and Woods, J. A Quantum Logic of Down Below. In: Engesser K, Gabbay D, Lehmann D, editors. Handbook of Quantum Logic and Quantum Structures. vol. 2; 2009.

[Buitelaar & Cimiano, 2008] Buitelaar, P., and Cimiano, P. (2008). Ontology learning and population: bridging the gap between text and knowledge, Vol. 167, Ios Press.

[Busemeyer & Bruza, 2012] Busemeyer, J., and Bruza, P. D. (2012). Quantum models of cognition and decision. Cambridge University Press.

[Camerer & Weber, 1992] Camerer, C., and Weber, M. Recent Developments in Modeling Preferences: Uncertainty and Ambiguity. Journal of Risk and Uncertainty. 1992; 5(4):325-370.

[Carnap, 1947] Carnap, R. Meaning and Necessity: A Study in Semantics and Modal Logic. University Of Chicago Press, Chicago, IL, USA (1947)

[Chang et al., 2013] Chang, Y., Diaz, F., Dong, A., Dumais, S., Radinsky, K., and Shokouhi, M. 2013. Temporal web dynamics and its application to information retrieval. WSDM 2013 tutorial.

[Charnov, 1976] Charnov, E.L. Optimal Foraging, the Marginal Value Theorem. Theoretical Population Biology. 1976; 9(2):129-136.

[Chi et al., 2001] Chi, E.H., Pirolli, P., Chen, K., and Pitkow, J. Using Information Scent to Model User Information Needs and Actions and the Web. In: Proceedings of CHI-01, 19th Conference on Human Factors in Computing Systems; 2001. p. 490-497.

[Christakis & Fowler, 2007] Christakis, N. A. and Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. New England Journal of Medicine, 357(4):370-379.

[Church et al., 2012] Church, G. M., Gao, Y., and Kosuri, S. (2012). Next-generation digital information storage in DNA. Science 337, p. 1628.

[Cimiano & Völker, 2005] Cimiano, P., and Völker, J. (2005). Text2Onto: A Framework for Ontology Learning and Data-driven Change Discovery. In International Conference on Application of Natural Language to Information Systems. Springer Berlin Heidelberg, pp. 227-238.

[Clark & Pulman, 2007] Clark, S., and Pulman, S. 2007. Combining Symbolic and Distributional Models of Meaning. Quantum Interaction, Papers from the 2007 AAAI Spring Symposium, Technical Report SS-07-08, Stanford, California, USA, March 26-28, 2007.

[Clark et al., 2013] Clark, S., Coecke, B., Grefenstette, E., Pulman, S., and Sadrzadeh, M. 2013. A quantum teleportation inspired algorithm produces sentence meaning from word meaning and grammatical structure. At http://arxiv.org/abs/1305.0556

[Coecke et al., 2010] Coecke, B., Sadrzadeh, M., and Clark, S. 2010. Mathematical Foundations for a Compositional Distributional Model of Meaning, at http://arxiv.org/abs/1003.4394

[Coecke et al., 2013] Coecke, B., Grefenstette, E., Sadrzadeh, M. 2013. Lambek vs. Lambek: Functorial vector space semantics and string diagrams for Lambek calculus. Annals of Pure and Applied Logic 164 (11), 1079-1100.

[Cohen & Widdows, 2015] Cohen, T., and Widdows, D. 2015. Embedding Probabilities in Predication Space with Hermitian Holographic Reduced Representations. In Proceedings of QI-15, Volume 9535 of the series Lecture Notes in Computer Science pp 245-257.

[Cohen et al., 2007] Cohen, J.D., McClure, S.M., and Yu, A.J. Should I Stay or Should I Go? How the Human Brain Manages the Trade-off between Exploitation and Exploration. Philosophical Transactions of the Royal Society B: Biological Sciences. 2007; 362(1481):933-942.

[Cohen et al., 2010] Cohen, T., Widdows, D., Schvaneveldt, R., and Rindflesch, T. (2010). Logical leaps and quantum connectives: Forging paths through predication space. In Proceedings of QI-10, 4th Symposium on Quantum Informatics for Cognitive, Social, and Semantic Processes, pages 11-13.

[Cole et al., 2010] Cole, M.J., Gwizdka, J., Bierig, R., Belkin, N.J., Liu, J., Liu, C., et al. Linking Search Tasks with Low-Level Eye Movement Patterns. In: Proceedings of ECCE-10, 28th European Conference on Cognitive Ergonomics; 2010. p. 109-116.

[Cole et al., 2013] Cole, M.J., Gwizdka, J., Liu, C., Belkin, N.J., snd Zhang, X. Inferring User Knowledge Level from Eye Movement Patterns. Information Processing & Management. 2013; 49(5):1075-1091.

[Cormack et al., 2009] Cormack, G.V., Clarke, C.L.A., and Buettcher, S. 2009. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR '09), 758-759.

[Cronin & Overfelt, 1994] Cronin, B., and Overfelt, K. (1994). The scholar's courtesy: A survey of acknowledgement behaviour. Journal of Documentation, 50(3):165-196.

[Cronin & Sugimoto, 2014] Cronin, B., and Sugimoto, C. R. (2014). Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact. MIT Press.

[Cutrell & Guan, 2007] Cutrell, E., and Guan, Z. What Are You Looking for?: An Eye-Tracking Study of Information Usage in Web Search. In: Proceedings of CHI-07, 25th Conference on Human Factors in Computing Systems. vol. 25; 2007. p. 407-416.

[Daránya & Eklund, 2007] Daránya, S., and Eklund, J. 2007. Automated text categorization of bibliographic records. Svensk biblioteksforskning/Swedish Library Research 16(2), 1-14.

[Daránya & Wittek, 2012] Daránya, S., and Wittek, P. (2012). Connecting the dots: Mass, energy, word meaning, and particle-wave duality. In Proceedings of QI-12, 6th International Quantum Interaction Symposium, pages 207-217.

[Daránya et al., 2016] Daránya, S., Wittek, P., Konstantinidis, K., Papadopoulos, S., and Kontopoulos, E. 2016. Physics as a metaphor to study semantic drift. In Proceedings of Semantics 2016 (forthcoming). See also at http://arxiv.org/abs/1608.01298 (16-08-04)

[de Jong et al., 2005] de Jong, F., Rode, H., and Hiemstra, D. 2005. Temporal Language Models for the Disclosure of Historical Text. In Proceedings of the XVIth International Conference of the Association for History and Computing, 161-168.

[de Solla Price, 1965] de Solla Price, D. J. (1965). Networks of scientific papers. Science, 149(3683):510-515.

[De Vine & Bruza, 2010] De Vine, L., and Bruza, P.D. (2010). Semantic oscillations: encoding context and structure in complex valued holographic vectors. In Quantum Informatics for Cognitive, Social, and Semantic Processes (QI 2010), AAAI Press, Arlington, Virginia.

[Deerwester et al., 1990] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., and Harshman, R. 1990. Indexing by latent semantic analysis. Journal of the American Society for Information Science 41, 6, 391-407.

[Di Buccio & Di Nunzio, 2011] Di Buccio, E., and Di Nunzio, G.: Envisioning dynamic quantum clustering in information retrieval. In: Proceedings of QI-11, 5th International Quantum Interaction Symposium, Aberdeen, UK (2011) 211-216.

[Ding & Peng, 2004] Ding, Z., & Peng, Y. (2004, January). A probabilistic extension to ontology language OWL. In *System Sciences, 2004. Proceedings of the 37th Annual Hawaii international conference on*(pp. 10-pp). IEEE.

[Dolby et al., 2009] Dolby, J., Fokoue, A., Kalyanpur, A., Schonberg, E. and Srinivas, K. (2009). Scalable highly expressive reasoner (SHER). In Web Semantics: Science, Services and Agents on the World Wide Web Journal, Vol. 7, Issue 4, pp. 357-361.

[Dominich, 2001] Dominich S. Mathematical Foundations of Information Retrieval. Kluwer Academic Publishers Norwell, MA, USA; 2001.

[Dowty et al., 1981] Dowty, D.R., Wall, R.E., and Peters, S. 1981. Introduction to Montague semantics. Springer.

[Dubinko et al., 2006] Dubinko, M., Kumar, R., Magnani, J., Kovak, J., Raghavan, P., and Tomkins, A. 2006. Visualizing tags over time. In Proceedings of the 15th International Conference on WWW, Scotland. 193-202.

[Dumais et al., 2010] Dumais, .S.T, Buscher, G., and Cutrell, E. Individual Differences in Gaze Patterns for Web Search. In: Proceeding of IIiX-10, 3rd Symposium on Information Interaction in Context. vol. 3; 2010. p. 185-194.

[Einstein et al., 1935] Einstein, A., Podolsky, B., and Rosen, N. (1935). Can quantum-mechanical description of physical reality be considered complete? Physical Review, 47(10):777.

[Eklund, 2016] Eklund, J. 2016. With or without context: Automatic text categorization using semantic kernels. Doctoral thesis. University of Borås.

[Ellsberg, 1961] Ellsberg D. Risk, Ambiguity, and the Savage Axioms. The Quarterly Journal of Economics. 1961;75(4):643-669.

[Engen et al., 2015] Engen, V., Veres, G., Crowle, S., Bashevoy, M., Walland, P., & Hall-May, M. (2015). A Semantic Risk Management Framework for Digital Audio-Visual Media Preservation.

[Falcao, 2010] Falcao, P. (2010). Developing a Risk Assessment Tool for the Conservation of Software-based Artworks (Doctoral dissertation, HKB).

[Fellbaum, 1998] Fellbaum, Ch. (Ed.) (1998). WordNet: An Electronic Lexical Database. MIT Press, Cambridge MA.

[Fisher, 1935] Fisher, R.A. The Design of Experiments. 9th Ed. New York, NY, USA: Hafner Press; 1935.

[Folland & Sitaram, 1997] Folland, G.B., and Sitaram, A. The Uncertainty Principle: A Mathematical Survey. Journal of Fourier Analysis and Applications. 1997; 3(3):207-238.

[Frey & Dueck, 2007] Frey, B. J., and Dueck, D. 2007. Clustering by passing messages between data points, Science 315(5814), 972-976.

[Frommholz et al., 2010] Frommholz, I., Larsen, B., Piwowarski, B., Lalmas, M., Ingwersen, P., and van Rijsbergen, K. 2010. Supporting polyrepresentation in a quantum-inspired geometrical retrieval framework. In Proceedings of the 3rd symposium on information interaction in context (IIiX '10), 115-124.

[Fyshe et al., 2013] Fyshe, A., Talukdar, P., Murphy, B., and Mitchell, T. 2013 Documents and Dependencies: an Exploration of Vector Space Models for Semantic Composition. In Proceedings of 17th Conference on Computational Natural Language Learning (CoNLL 2013).

[Gangemi et al., 2002] Gangemi, Aldo et al. “Sweetening Ontologies with DOLCE.” Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web, Lecture Notes in Computer Science, vol. 2473 (2002): 223–233.

[Garfield et al., 1964] Garfield, E., Sher, I., and Torpie, R. (1964). The use of citation data in writing the history of science. Report 99, The Institute for Scientific Information.

[Garfield et al., 2003] Garfield, E., Pudovkin, A. I., and Istomin, V. S. (2003). Why do we need algorithmic historiography? Journal of the American Society for Information Science and Technology, 54(5):400-412.

[Garfield, 2009] Garfield, E. (2009). From the science of science to scientometrics visualizing the history of science with HistCite software. Journal of Informetrics, 3(3):173-179.

[Gilbert, 1977] Gilbert, N. G. (1977). Referencing as persuasion. Social Studies of Science, 7(1):113-122.

[Giugno & Lukasiewicz, 2002] Giugno, R., & Lukasiewicz, T. (2002, September). P-SHOQ (D): A probabilistic extension of SHOQ (D) for probabilistic ontologies in the semantic web. In *JELIA* (Vol. 2, pp. 86-97).

[Goldberg & Kotval, 1999] Goldberg, J.H., and Kotval, X.P. Computer Interface Evaluation Using Eye Movements: Methods and Constructs. International Journal of Industrial Ergonomics. 1999 Oct; 24(6):631-645.

[Goldman et al., 2013] Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProus, E.M., Sipos, B., and Birney, E. (2013). Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature 494, pp. 77–80.

[Grass et al., 2015] Grass, R.N., Heckel, R., Puddu, M., Paunescu, D., and Stark, W.J. (2015). Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes. Angewandte Chemie International Edition 54, pp. 2552 –2555.

[Grefenstette, 2013] Grefenstette, E. 2013. Category-theoretic quantitative compositional distributional models of natural language semantics. PhD thesis, Oxford University.

[Guarino & Welty, 2000] Guarino, Nicola, and Christopher Welty. "A Formal Ontology of Properties." Knowledge Engineering and Knowledge Management Methods, Models, and Tools. Vol. 1937. Springer Berlin Heidelberg, 2000. 97–112.

[Gulla et al., 2010] Gulla, J. A., Solskinnsbakk, G., Myrseth, P., Haderlein, V., and Cerrato, O. 2010. Semantic Drift in Ontologies. In WEBIST 2, 13-20.

[Gwidzka, 2014] Gwizdka J. Characterizing Relevance with Eye-tracking Measures. In: Proceedings of IIiX-14, 5th Information Interaction in Context Symposium. vol. 5; 2014. p. 58-67.

[Haarslev & Möller, 2003] Haarslev, V., and Möller, R. (2003, October). Racer: A Core Inference Engine for the Semantic Web. In EON (Vol. 87).

[Harris, 1968] Harris, Z. 1968. Mathematical structures of language. Interscience Publishers.

[Harris, 1970] Harris, Z.: Distributional structure. In Harris, Z., ed.: Papers in structural and transformational Linguistics. Formal Linguistics. Humanities Press, New York, NY, USA (1970) 775-794

[Haven, 2015] Haven, E. (2015). Financial payoff functions and potentials. In Proceedings of QI-14, 8th International Conference on Quantum Interaction, pages 189-195.

[Heisenberg, 1927] Heisenberg, W. Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. Zeitschrift für Physik. 1927 Mar; 43(3-4):172-198.

[Hersh et al., 1994] Hersh, W., Buckley, C., Leone, T.J., and Hickam, D. OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research. In: Proceedings of SIGIR-94, 17th International Conference on Research and Development in Information Retrieval. vol. 17; 1994. p. 192-201.

[Hertwig & Erev, 2009] Hertwig R, Erev I. The Description-Experience Gap in Risky Choice. Trends in Cognitive Sciences. 2009; 13(12):517-523.

[Heunen et al., 2013] Heunen, C., Sadrzadeh, M., and Grefenstette, E. (Eds.) Quantum physics and linguistics: a compositional, diagrammatical discourse. Oxford University Press.

[Hicks, 2012] Hicks, D. (2012). Performance-based university research funding systems. Research Policy, 41(2):251-261.

[Hiley & Pilkkänen, 1977] Hiley, B., and Pylkkänen, P. (1993). Active information and cognitive science - A reply to Kieseppä. In Pylkkänen, P., Pylkkö, P., and Hautamäki, A. (Eds.): Brain, Mind and Physics. IOS Press, Amsterdam, pp. 123-145.

[Hills et al, 2015] Hills, T.T., Todd, P.M., Lazer, D., Redish, A.D., Couzin, I.D. Exploration versus Exploitation in Space, Mind, and Society. Trends in Cognitive Sciences. 2015 Jan; 19(1):46-54.

[Hills et al., 2012] Hills, T.T., Jones, M.N., and Todd, P.M. Optimal Foraging in Semantic Memory. Psychological Review. 2012; 2:431-440.

[Hirsch & Quapp, 2004] Hirsch, M., and Quapp, W. 2004. The reaction pathway of a potential energy surface as curve with induced tangent. Chemical Physics Letters 395, 150-156.

[Hochreiter & Schmidhuber, 1997] Hochreiter, S., and Schmidhuber, J. 1997. Long Short-Term Memory. Neural Computation 9, 8, 1735-1780.

[Horn & Gottlieb, 2001] Horn, D. and Gottlieb, A. (2001). Algorithm for Data Clustering in Pattern Recognition Problems Based on Quantum Mechanics. Physical Review Letters 88(1), 018702.

[Horn & Gottlieb, 2001] Horn, D., and Gottlieb, A.: The method of quantum clustering. In Dietterich, T., Becker, S., Ghahramani, Z., eds.: Advances in Neural Information Processing Systems. Volume 14, Vancouver, Canada (2001) 769-776.

[Horridge et al., 2006] Horridge, M., Drummond, N., Goodwin, J., Rector, A., Stevens, R., and Wang, H. H. (2006). The Manchester OWL Syntax. OWL Experiences and Directions Workshop (OWLED) 2006, Vol. 216.

[Horrocks et al., 2004] Horrocks, I., Patel-Schneider, P. F., Boley, H., Tabet, S., Grosof, B., and Dean, M. (2004). SWRL: A semantic web rule language combining OWL and RuleML. W3C Member submission, 21, 79.

[Horty, 2001] Horty, J. F., 2001, "Nonmonotonic Logic," in Goble, Lou, ed., The Blackwell Guide to Philosophical Logic. Blackwell.

[ICA, 2000] ICA - International Council on Archives (2000), ISAD(G): General International Standard Archival Description, Second Edition. At: http://www.ica.org/10207/standards/isadg-general-international-standard-archival-description-second-edition.html

[Jaccard, 1912] Jaccard, P. 1912. The distribution of the flora in the alpine zone. New Phytologist 11(2), 37-50.

[Jantzen, 2007] Jan Jantzen, *Foundations of Fuzzy Control*. Wiley, 2007.

[Jarrett, 2004] Jarrett, T. 2004. Large Scale Structure in the Local Universe: The 2MASS Galaxy Catalog. At http://xxx.lanl.gov/html/astro-ph/0405069v1

[Kammerer & Gerjects, 2012] Kammerer, Y., and Gerjets, P. Effects of Search Interface and Internet-Specific Epistemic Beliefs on Source Evaluations during Web Search for Medical Information: An Eye-Tracking Study. Behaviour & Information Technology. 2012; 31(1):83-97.

[Kanerva et al., 2010] Kanerva, P., Kristofersson, J., and Holst, A. 2000. Random indexing of text samples for latent semantic analysis. In Proceedings of the 22nd Annual Conference of the Cognitive Science Society, vol. 1036.

[Kaplan, 1965] Kaplan, N. (1965). The norms of citation behavior: Prolegomena to the footnote. American Documentation, 16(3):179-184.

[Kazansky et al., 2016] Kazansky, P., Cerkauskaite, A., Beresna, M., Drevinskas, R., Patel, A., Zhang, J., and Gecevicius, M. (2016). Eternal 5D data storage via ultrafast-laser writing in glass. SPIE Optoelectronics & Communications. February 17, 2016.

[Ke et al., 2015] Ke, Q., Ferrara, E., Radicchi, F., and Flammini, A. (2015). Defining and identifying sleeping beauties in science. Proceedings of the National Academy of Sciences, 112(24):7426-7431.

[Kemman et al., 2013] Kemman, M., Kleppe, M., and Maarseveen, J. Eye Tracking the Use of a Collapsible Facets Panel in a Search Interface. In: Aalberg T, Papatheodorou C, Dobreva M, Tsakonas G, Farrugia C, editors. Research and Advanced Technology for Digital Libraries: Proceedings of TPDL-13, 17th International Conference on Theory and Practice of Digital Libraries. vol. 8092; 2013. p. 405-408.

[Kessler, 1963] Kessler, M. M. (1963). Bibliographic coupling between scientific papers. American Documentation, 14(1):10-25.

[Khrennikov, 2010] Khrennikov, A. (2010). Ubiquitous Quantum Structure: From Psychology to Finance. Springer Verlag.

[Kim et al., 2016] Kim J, Thomas, P., Sankaranarayana, R., Gedeon, T., and Yoon H.J. Understanding Eye Movements on Mobile Devices for Better Presentation of Search Results. Journal of the Association for Information Science & Technology. 2016.

[Klein & Fensel, 2001] Klein, M. C., and Fensel, D. 2001. Ontology versioning on the Semantic Web. In Proceedings of SWWS, 75-91.

[Knight, 1921] Knight FH. Risk, Uncertainty and Profit. Houghton Miffin, Boston; 1921.

[Knublauch et al., 2011] Knublauch, H., Hendler, J. A., and Idehen, K. (2011). SPIN - overview and motivation. World Wide Web Consortium, W3C Member Submission, available online: https://www.w3.org/Submission/spin-overview/

[Kohonen, 2001] Kohonen, T. 2001. Self-Organizing Maps. Springer.

[Kolling et al., 2012] Kolling, N., Behrens, T.E.J., Mars, R.B., and Rushworth, M.F.S. Neural Mechanisms of Foraging. Science. 2012; 336(6077):95-98.

[Kontopoulos et al., 2012] E. Kontopoulos, T. Zetta, and N. Bassiliades, "Semanti-cally-enhanced authoring of defeasible logic rule bases in the semantic web," in Proceedings of the 2nd Inter-national Conference on Web Intelligence, Mining and Semantics, 2012, p. 56.

[Kontopoulos et al., 2016] Kontopoulos, E., Riga, M., Mitzias, P., Andreadis, S., Stavropoulos, T. G., Lagos, N., Vion-Dury, J.-Y., Meditskos, G., Falcão, P., Laurenson, P., and Kompatsiaris, I. (2016) Ontology-based Representation of Context of Use in Digital Preservation. In 1st Workshop on Humanities in the Semantic Web (WHiSe), CEUR Workshop Proceedings, Vol. 1608, pp. 65-72.

[Koolen et al., 2009] Koolen, M., Kamps, J., and de Keijzer, V. 2009. Information Retrieval in Cultural Heritage. Interdisciplinary Science Reviews 34, 2-3, 268-284.

[Kules et al., 2009] Kules, B., Capra, R., Banta, M., and Sierra, T. What Do Exploratory Searchers Look at in a Faceted Search Interface? In: Proceedings of JCDL-09, 9th Joint Conference on Digital Libraries; 2009. p. 313-322.

[Lagos et al., 2016] Lagos, N., Riga, M., Mitzias, P., Vion-Dury, J.Y., Kontopoulos, E., Waddington, S., Laurenson, P., and Kompatsiaris, I. Dependency Modelling for Inconsistency Management in Digital Preservation – The PERICLES Approach. Submitted to Information Systems Frontiers, under revision, Springer.

[Lam & Governatori, 2009] H.-P. Lam and G. Governatori, "The making of SPIN-dle," in Rule Interchange and Applications, Springer, 2009, pp. 315–322.

[Lambek, 2011] Lambek, J. 2011. Compact monoidal categories from linguistics to physics. In Coecke, B. (Ed.) New structures for physics. Springer.

[Laserre, 2001] Lasserre, J. (2001). Global optimization with polynomials and the problem of moments. SIAM Journal on Optimization, 11(3):796-817.

[LeCun et al., 2008] LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F.J. 2008. A Tutorial on Energy-Based Learning. In Bakir, G., Hofmann, T., Schölkopf, B., Smola, A.J., Taskar, B., and Vishwanathan, S.V.N. (Eds.) Predicting structured data. MIT Press.

[Lorigo et al., 2008] Lorigo, .L, Haridasan, M., Brynjarsdottir, H., Xia, L., Joachims, T., Gay, G., et al. Eye Tracking and Online Search: Lessons Learned and Challenges Ahead. Journal of the American Society for Information Science & Technology. 2008; 59(7):1041-1052.

[Lötsch & Ultsch, 2014] Lötsch, J., and Ultsch, A. 2014. Exploiting the structures of the U-matrix. In Advances in Self-Organizing Maps and Learning Vector Quantization: Proceedings of WSOM, Mittweida, Germany, July 2-4, 249-258.

[Lukasiewicz, 2008] Lukasiewicz, T. (2008). Expressive probabilistic description logics. *Artificial Intelligence*, *172*(6), 852-883.

[Ma et al., 2015] Ma, L., Krishnan, R., and Montgomery, A.L. 2015. Latent homophily or social influence? An empirical analysis of purchase within a social network. Management Science, 61(2):454-473.

[Maarala et al., 2016] Maarala, A. I., Su, X., & Riekki, J. (2016). Semantic Reasoning for Context-aware Internet of Things Applications. arXiv preprint arXiv:1604.08340.

[Mantegna & Stanley, 2000] Mantegna, R. N. and Stanley, H. E. (2000). Introduction to Econophysics. Cambridge University Press, Cambridge.

[McArthur & Pianka, 1966] MacArthur, R.H., and Pianka, E.R. On Optimal Use of a Patchy Environment. American Naturalist. 1966; 100(916):603-609.

[McCain, 1985] McCain, K. W. (1986). Cocited author mapping as a valid representation of intellectual structure. Journal of the American Society of Information Science, 37(3):111-122.

[McGuinness & Da Silva, 2004] McGuinness, D. L., & Da Silva, P. P. (2004). Explaining answers from the semantic web: The inference web approach. Web Semantics: Science, Services and Agents on the World Wide Web, 1(4), 397-413.

[McIlroy & McLevey, 2015] McIlroy-Young, R. and McLevey, J. (2015). Metaknowledge: open source software for social networks, bibliometrics, and sociology of knowledge research. Waterloo, ON, Canada.

[Meroño-Peñuela & Hoekstra, 2014] Meroño-Peñuela, Albert, and Rinke Hoekstra. "What Is Linked Historical Data?" Springer International Publishing, 2014. 282–287.

[Meroño-Peñuela et al., 2013] Meroño-Peñuela, A., Guéret, C., Hoekstra, R., and Schlobach, S. 2013. Detecting and reporting extensional concept drift in statistical linked data. In Proceedings of the 1st International Workshop on Semantic Statistics (SemStats 2013), ISWC 2013.

[Mezey, 1999] Mezey, P. 1999. The topology of catchment regions of potential energy hypersurfaces. Theoretical Chemistry Accounts 102(1), 279-284.

[Mihalcea & Moldovan, 1998] Mihalcea, R., and Moldovan, D.I. Word sense disambiguation based on semantic density. In: Proceedings of COLING-ACL, 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. (1998)

[Mikolov et al., 2013] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J. 2013. Distributed Representations of Words and Phrases and their Compositionality. At https://arxiv.org/abs/1310.4546.

[Milojevic, 2015] Milojevic, S. 2015. Quantifying the Cognitive Extent of Science. At http://arxiv.org/abs/1511.00040

[Mitroff, 1974] Mitroff, I. I. (1974). Norms and counter-norms in a select group of the Apollo Moon scientists: A case study of the ambivalence of scientists. American Sociological Review, 39(4):579.

[Mitzias et al., 2015] Mitzias, P., Riga, M., Waddington, S., Kontopoulos, E., Meditskos, G., Laurenson, P., and Kompatsiaris, I. (2015). An Ontology Design Pattern for Digital Video. Proc. 6th Workshop on Ontology and Semantic Web Patterns (WOP 2015) co-located with 14th Int. Semantic Web Conf. (ISWC 2015), CEUR-WS Vol-1461, Bethlehem, Pennsylvania, USA, October 11-15.

[Mitzias et al., 2016] Mitzias, P., Riga, M., Kontopoulos, E., Stavropoulos, T. G., Andreadis, S., Meditskos, G. and Kompatsiaris, I. (2016). User-driven Ontology Population from Linked Data Sources. In International Conference on Knowledge Engineering and Semantic Web, Prague, Czech Republic, 21-23 Sept., 2016 (accepted for publication).

[Moschitti, 2010] Moschitti, A. 2010. Kernel Engineering for Fast and Easy Design of Natural Language Applications. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10), 1-91.

[Motik et al., 2009] Motik, B., Shearer, R., & Horrocks, I. (2009). Hypertableau reasoning for description logics. Journal of Artificial Intelligence Research, 36(1), 165-228.

[Mugur-Schächter, 2014] Mugur-Schächter, M. (2014). On the concept of probability. Mathematical Structures in Computer Science, 24(03):e240309.

[Navascués et al., 2007] Navascués, M., Pironio, S., and Acín, A. (2007). Bounding the set of quantum correlations. Physical Review Letters, 98(1):10401.

[Nielsen & Chuang, 2000] Nielsen, M.A., and Chuang, I.L. 2000. Quantum computation and quantum information. Cambridge University Press.

[Nute, 1994] Nute, D. (1994). Defeasible logic. Handbook of Logic in Artificial Intelligence and Logic Programming, 3, 353–395.

[Osgood et al., 1957] Osgood, C.E., Suci, G.J., and Tannenbaum, P.H. 1957. The Measurement of Meaning. University of Illinois Press: Urbana.

[Page & Brin, 1998] Page, L., and Brin, S. 1998. The PageRank Citation Ranking: Bringing Order to the Web. At http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

[Pareti et al., 2015] Pareti, P., Klein, E., and Barker, A. D. 2015. A Linked Data scalability challenge: concept reuse leads to semantic decay. In Proceedings of WebSci'15 ACM Web Science Conference.

[Pearl, 2009] Pearl, J. (2009). Causality: Models, Reasoning and Inference. Cambridge University Press, New York, NY, USA.

[Peng et al., 2009] Peng, L., Yang, B., Chen, Y., and Abraham, A. 2009. Data gravitation based classification. Information Sciences 179, 809-819.

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C.D. 2014. GloVe: Global Vectors for Word Representation. In Proceedings of Empirical Methods in Natural Language Processing (EMNLP), 1532-1543.

[PERICLES D2.3.2, 2015] PERICLES Consortium, Deliverable 2.3.2: Data Surveys and Domain Ontologies, September 2015.

[PERICLES D3.3, 2015] PERICLES Consortium, Deliverable 3.3: Semantics for Change Management, July 2015.

[PERICLES D4.3, 2016] PERICLES Consortium, Deliverable 4.3: Content Semantics and Use Context Analysis Techniques, January 2016.

[PERICLES D4.4, 2016] PERICLES Consortium, Deliverable 4.4: Modelling Contextualised Semantics, May 2016.

[PERICLES D5.3, 2016] PERICLES Consortium, Deliverable 5.3: Complete Tool Suite for Ecosystem Management and Appraisal Processes, September 2016.

[Peterson et al., 2003] Peterson, E.R., Deary, I.J., and Austin, E.J. The Reliability of Riding's Cognitive Style Analysis Test. Personality and Individual Di_erences. 2003; 34:881- 891.

[Peterson, 2005] Peterson, E.R. Verbal Imagery Cognitive Styles Test & Extended Cognitive Style Analysis-Wholistic Analytic Test Administration Guide; 2005.

[Pietras et al., 2003] Pietras, C.J, Locey, M.L, and Hackenberg, T.D. Human Risky Choice under Temporal Constraints: Tests of an Energy-Budget Model. Journal of the Experimental Analysis of Behavior. 2003; 80(1):59.

[Pirolli & Card, 1999] Pirolli, P., and Card, S. Information Foraging. Psychological Review. 1999; 106(4):643.

[Pirolli et al., 2000] Pirolli, P., Card, S.K., and Van Der Wege, M.M. The Effect of Information Scent on Searching Information: Visualizations of Large Tree Structures. In: Proceedings of the Working Conference on Advanced Visual Interfaces; 2000. p. 161-172.

[Pirolli et al., 2001] Pirolli, P., Card, S.K, and Van Der Wege, M.M. Visual Information Foraging in a Focus + Context Visualization. In: Proceedings of CHI-01, 19th Conference on Human Factors in Computing Systems; 2001. p. 506-513.

[Pironio et al., 2010] Pironio, S., Navascués, M., and Acín, A. (2010). Convergent relaxations of polynomial optimization problems with noncommuting variables. SIAM Journal on Optimization, 20(5):2157-2180.

[Plate, 1991] Plate, T. 1991. Holographic Reduced Representations: Convolution Algebra for Compositional Distributed Representations. In Proceedings of International Joint Conference of Artificial Intelligence, pp. 30-35.

[Prud'Hommeaux & Seaborne, 2008] Prud'Hommeaux, E., and Seaborne, A. (2008). SPARQL query language for RDF. W3C recommendation, 15.

[Pulman, 2013] Pulman, S. 2013. Distributional semantic models. In Heunen, C., Sadrzadeh, M., and Grefenstette, E. (Eds.) Quantum physics and linguistics: a compositional, diagrammatical discourse. Oxford University Press. 333-358.

[Ramirez & Steudel, 2008] Ramirez, Y.W., and Steudel, H.J. 2008. Measuring knowledge work: the knowledge work quantification framework. Journal of Intellectual Capital 9(4), pp. 564-584.

[Rice, 2015] Rice, D. (2015). Sustaining Consistent Video Presentation. Tate Research Articles, March.

[Robertson & Spärck Jones, 1976] Robertson, S.E., and Spärck Jones, K. 1976. Relevance weighting of search terms. JASIS, 27, 3, 129-146.

[Robertson & Zaragoza, 2009] Robertson, S., and Zaragoza, H. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval 3, 4, 333-389.

[Rohaninezhad et al., 2015] Rohaninezhad, M., Arif, S. M., & Noah, S. A. M. (2015). A grounder for SPINdle defeasible logic reasoner. *Expert Systems with Applications*, *42*(20), 7098-7109.

[Rosengren, 1968] Rosengren, K. E. (1968). Sociological Aspects of the Literary System. Natur och Kultur, Stockholm.

[Rushworth et al., 2012] Rushworth, M.F., Kolling, N., Sallet, J., and Mars, R.B. Valuation and Decision-Making in Frontal Cortex: One or Many Serial or Parallel Systems? Current Opinion in Neurobiology. 2012 Dec; 22(6):946-955.

[Sadrzadeh & Grefenstette, 2011] Sadrzadeh, M., and Grefenstette, E. 2011. A Compositional Distributional Semantics, Two Concrete Constructions, and Some Experimental Evaluations. In Proceedings of QI -11, 35-47.

[Salton, 1975] Salton, G. 1975. Dynamic information and library processing. Prentice-Hall: Upper Saddle River.

[Sandstrom, 2001] Sandstrom, P. (2001). Scholarly communication as a socioecological system. Scientometrics, 51(3):573-605.

[Savenkov et al., 2011] Savenkov, D., Braslavski, P., and Lebedev, M. Search Snippet Evaluation at Yandex: Lessons Learned and Future Directions. In: Forner, P., Gonzalo, J., Kekäläinen, J., Lalmas, M., and de Rijke, M., editors. Multilingual and Multimodal Information Access Evaluation. vol. 6941. Heidelberg, Germany: Springer; 2011. p. 14-25.

[Schlieder, 2010] Schlieder, C. 2010. Digital heritage: Semantic challenges of long-term preservation. Semantic Web 1, 1-2, 143-147.

[Schölkopf & Schmola, 2002], Schölkopf, B., and Smola, A. J. (2002). Learning with kernels: support vector machines, regularization, optimization, and beyond. Cambridge, Mass.: MIT Press.

[Schütze, 1998] Schütze, H. 1998. Automatic word sense discrimination. Computational Linguistics 24(1), 97-123.

[Shaparenko et al., 2005] Shaparenko, B., Caruana, R., Gehrke, J., and Joachims, T. 2005. Identifying temporal patterns and key players in document collections. In Proceedings of ICDM, 165-174.

[Shearer et al., 2008] Shearer, R., Motik, B., and Horrocks, I. (2008). HermiT: A Highly-Efficient OWL Reasoner. In OWLED, Vol. 432, p. 91.

[Sirin et al., 2007] Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., and Katz, Y. (2007). Pellet: A practical owl-dl reasoner. Web Semantics: science, services and agents on the World Wide Web, 5(2), pp. 51-53.

[Small, 1973] Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society of Information Science, 24(4):265-269.

[Szabó, 2013] Szabó, Z. G. 2013. "Compositionality": The Stanford Encyclopedia of Philosophy (Fall 2013 Edition), Edward N. Zalta (ed.), at http://plato.stanford.edu/archives/fall2013/entries/compositionality/

[Taira et al., 2007] Taira, R.K., Bashyam, V., and Kangarloo, H. 2007. A Field Theoretical Approach to Medical Natural Language Processing. IEEE Transactions on Information Technology in Biomedicine 11(4), 364-375.

[Tanaka & Wang, 2001] Kazuo Tanaka; Hua O. Wang (2001). *Fuzzy control systems design and analysis: a linear matrix inequality approach*. John Wiley and Sons.

[Tosi et al., 2014] Tosi, A., Olier, I., and Vellido, A. 2014. Probability ridges and distortion flows: Visualizing multivariate time series using a variational Bayesian manifold learning method. In Advances in Self-Organizing Maps and Learning Vector Quantization: Proceedings of WSOM, Mittweida, Germany, July 2-4. 55-64.

[Tozzi & Peters, 2016] Tozzi, A., and Peters, J.F. (2016). Towards a fourth spatial dimension of brain activity. Cognitive Neurodynamics 10(3), pp. 189-199. DOI: 10.1007/s11571-016-9379-z

[Trier, 1934] Trier, J. 1934. Das sprachliche Feld. Neue Jahrbücher für Wissenschaft und Jugendbildung, 10, 428-449.

[Tsarkov & Horrocks, 2006] Tsarkov, D., and Horrocks, I. (2006). FaCT++ description logic reasoner: System description. In Automated reasoning (pp. 292-297). Springer Berlin Heidelberg.

[Turney & Pantel, 2010] Turney, P.D., and Pantel, P. (2010) From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research 37, 1, 141-188.

[Tury & Bieliková, 2006] Tury, M., and Bieliková, M. 2006. An approach to detection of ontology changes. In Workshop proceedings of the sixth ACM international conference on Web engineering, 14.

[Uexküll & Kriszat, 1956] Uexküll, J.J. and Kriszat, G. 1956. Streifzüge durch die Umwelten von Tieren und Menschen: Ein Bilderbuch unsichtbarer Welten. Bedeutungslehre. Rowohlt.

[Ultsch, 2005] Ultsch, A. 2005. Clustering with SOM: U* C. In Proceedings of the 5th Workshop on Self-Organizing Maps, 2, 75-82.

[Uschold, 2006] Uschold, M. (2000). Creating, integrating and maintaining local and global ontologies. In Proceedings of the First Workshop on Ontology Learning (OL-2000) in conjunction with the 14th European Conference on Artificial Intelligence (ECAI-2000). Citeseer.

[Vakkari et al., 2014] Vakkari, P., Luoma, A., Pöntinen, J. Books' Interest Grading and Dwell Time in Metadata in Selecting Fiction. In: Proceedings of IIiX-14, 5th Information Interaction in Context Symposium; 2014. p. 28-37.

[van Eck & Waltman, 2010] van Eck, N. J. and Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84(2):523-538.

[van Eck & Waltman, 2014] van Eck, N. J. and Waltman, L. (2014). Visualizing bibliometric networks. In Measuring Scholarly Impact, pages 285-320. Springer Science + Business Media.

[van Rijsbergen, 2004] van Rijsbergen, C.J. The Geometry of Information Retrieval. New York, NY, USA: Cambridge University Press; 2004.

[Vanclay, 2012] Vanclay, J. K. (2012). Impact factor: outdated artefact or stepping-stone to journal certification? Scientometrics, 92(2):211-238.

[Vapnik, 1998] Vapnik, V. 1998. Statistical learning theory. New York: Wiley.

[Velardi et al., 2005] Velardi, P., Navigli, R., Cucchiarelli, A. and Neri, F. (2005). Evaluation of OntoLearn, a methodology for automatic learning of domain ontologies. In Ontology Learning from Text: Methods, evaluation and applications, 123, 92, IOS Press.

[Veltman, 1996] Veltman, F. 1996. Defaults in update semantics. Journal of Philosophical Logic 25, 3, 221-261.

[Ver Steeg & Galstyan, 2011] Ver Steeg, G. and Galstyan, A. (2011). A sequence of relaxations constraining hidden variable models. In Proceedings of UAI-11, 27th Conference on Uncertainty in Artificial Intelligence, pages 717-726.

[Ver Steeg, 2015] Ver Steeg, G. L. (2015). Bell inequalities for complex networks. Technical report, University of Southern California.

[Vion-Dury et al., 2015] Vion-Dury, J.Y et al. Designing for inconsistency–the dependency-based PERICLES approach. East European Conference on Advances in Databases and Information Systems. Springer International Publishing, 2015.

[W3C, 2012] OWL 2 Web Ontology Language Document Overview (Second Edition). W3C Recommendation 11 December 2012. At: http://www.w3.org/TR/owl2-overview/

[Wang & Gai, 2014] Wang, L., and Gai, S. (2014). The next generation mass storage devices – Physical principles and current status. Contemporary Physics, 55:2, pp. 75-93, DOI:10.1080/00107514.2013.878565.

[Wang et al., 2011] Wang, S., Schlobach, S., and Klein, M. 2011. Concept drift and how to identify it. Web Semantics: Science, Services and Agents on the World Wide Web, 9, 3, 247-265.

[Wang et al., 2011] Wang, Shenghui, Stefan Schlobach, and Michel Klein. “Concept Drift and How to Identify It.” Journal of Web Semantics 9.3 (2011).

[Webb et al., 2016] Webb, G.I., Hyde, R., Cao, H., Hai Long, N., and Petitjean, F. 2016. Characterizing concept drift. To appear in Data Mining and Knowledge Discovery. At arxiv.org/abs/1511.03816

[Weinstein & Horn, 2009] Weinstein, M., and Horn, D. Dynamic quantum clustering: A method for visual exploration of structures in data. Physical Review E 80(6) (2009) 066117.

[White & Griffith, 1981] White, H. D. and Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. Journal of the American Society of Information Science, 32(3):163-171.

[White, 2002] White, H. 2002. Cross-textual cohesion and coherence. In Proceedings of the Workshop on Discourse Architectures: The Design and Analysis of Computer-Mediated Conversation, Minneapolis, MN, USA.

[White, 2016a] White, R.W. (2016). Models and Frameworks for Information Seeking. In: Interactions with Search Systems:. Cambridge University Press, pp. 97–138.

[White, 2016b] White, R.W. Interactions with Search Systems. New York, NY, USA: Cambridge University Press; 2016.

[Widdows & Cohen, 2015] Widdows, D., and Cohen, T. 2014. Reasoning with Vectors: A Continuous Model for Fast Robust Inference. Logic Journal of the IGPL 23(2), 141-173.

[Williams et al., 2005] Williams, P., Li, S., Feng, J., and Wu, S. 2005. Scaling the Kernel Function to Improve Performance of the Support Vector Machine. In Wang, J., Liao, X. and Yi, Z. (Eds.), LNCS 3496, pp. 831–836. Springer: Berlin.

[Wittek & Darányi, 2011] Wittek, P., and Darányi, S. Spectral composition of semantic spaces. In: Proceedings of QI-11, 5th International Quantum Interaction Symposium. (2011)

[Wittek et al., 2013a] Wittek, P., Lim, I.S., and Rubio-Campillo, X. Quantum Probabilistic Description of Dealing with Risk and Ambiguity in Foraging Decisions. In: Proceedings of QI-13, 7th International Quantum Interaction Symposium; 2013. p. 296-307.

[Wittek et al., 2013b] Wittek, P., Koopman, B., Zuccon, G., and Darányi, S. 2013. Combining Word Semantics within Complex Hilbert Space for Information Retrieval. In Proceedings of QI-13.

[Wittek et al., 2014] Wittek, P., Darányi, S., and Liu, Y.H. 2014. A vector field approach to lexical semantics. Proceedings of QI-14.

[Wittek et al., 2015a] Wittek, P., Darányi, S., Kontopoulos, E., Moysiadis, T., and Kompatsiaris, I. 2015. Monitoring term drift based on semantic consistency in an evolving vector field. In International Joint Conference on Neural Networks (IJCNN), 2015, 1–8.

[Wittek et al., 2015b] Wittek, P., Gao, S. C., Lim, I. S., and Zhao, L. 2015. Somoclu: An efficient parallel library for self-organizing maps. At http://arxiv.org/pdf/1305.1422.pdf

[Wittek et al., 2016a] Wittek, P., Liu, Y.H., Darányi, S., Gedeon, T., and Lim, I.S. 2016. Risk and Ambiguity in Information Seeking: Eye Gaze Patterns Reveal Contextual Behaviour in Dealing with Uncertainty. https://arxiv.org/abs/1606.08157. Submitted to Frontiers in Psychology (16-07-24).

[Wittek et al., 2016b] Wittek, P., Darányi, S., and Nelhans, G. 2016. Ruling Out Static Latent Homophily in Citation Networks. At http://arxiv.org/pdf/1605.08185v1.pdf

[Wittek, 2014] Wittek, P. 2014. Quantum Machine Learning: What Quantum Computing Means to Data Mining. Elsevier: Amsterdam.

[Wittek, 2015] Wittek, P. 2015. Ncpol2sdpa: Sparse Semidefinite Programming Relaxations for Polynomial Optimization Problems of Noncommuting Variables. ACM Transactions on Mathematical Software 41(3), 21. At https://arxiv.org/abs/1308.6029

[Wittgenstein, 1963] Wittgenstein, L. 1963. Philosophical investigations. Blackwell: Oxford. 43.

[Wong et al., 1985] Wong, S. K. M., Ziarko, W., and Wong, P. C. N. 1985. Generalized vector spaces model in information retrieval. In Proceedings of ACM SIGIR'85, pp 18-25.

[Yildiz, 2006] Yildiz, B. (2006). Ontology evolution and versioning. Vienna University of Technology, Karlsplatz.

[Yukalov & Sornette, 2008] Yukalov, V.I., and Sornette, D. Quantum Decision Theory as Quantum Theory of Measurement. Physics Letters A. 2008 Nov; 372(46):6867-6871.

[Yuret, 1998] Yuret, D. 1998. Discovery of linguistic relations using lexical attraction. PhD thesis. At http://www2.denizyuret.com/pub/yuretphd.pdf

[Zahedi et al., 2014] Zahedi, Z., Costas, R., and Wouters, P. (2014). How well developed are altmetrics? A cross-disciplinary analysis of the presence of "alternative metrics" in scientific publications. Scientometrics, 101(2):1491-1513.

[Zeng & Coecke, 2016] Zeng, W., and Coecke, B. 2016. Quantum Algorithms for Compositional Natural Language Processing. In Proceedings of SLPCS 2016. At https://arxiv.org/abs/1608.01406.

[Zhang et al., 2015] Zhang, X., Liu, J., Cole, M., and Belkin, N. Predicting Users' Domain Knowledge in Information Retrieval Using Multiple Regression Analysis of Search Behaviors. Journal of the Association for Information Science & Technology. 2015; 66(5):980-1000.

[Ziman, 2000] Ziman, J. (2000). Real Science. What it is, and what it means. Cambridge University Press.