GRANT AGREEMENT: 601138 | SCHEME FP7 ICT 2011.4.3
Promoting and Enhancing Reuse of Information throughout the Content Lifecycle taking account of Evolving Semantics [Digital Preservation]

Drift-a-LOD workshop, late breaking results session
@ EKAW 2016, Bologna, 20 November 2016

Physics as a Metaphor to Model Semantic Drift: Exploring the Tate Collection Metadata

Darányi, S.*, Wittek, P.*, Konstantinidis, K.**

- University of Borås, Sweden
- The Institute of Photonic Sciences (ICFO), Spain

** The Centre for Research & Technology Hellas (CERTH), Greece

SEVENTH FRAMEWORK PROGRAMME

Basic idea and past work on the topic

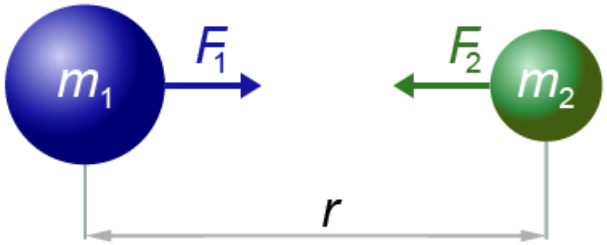
- Approach: word meaning is substance superimposed on form in continuous field (cf. Trier 1934)
  - Needs a vector field semantics
  - Assumes that the dynamics of such a field goes back to some “energy” stored in structures, i.e. word/sentence meaning is “energy”

- Past work – Theory:
  - Wittek, P., Daránvi, S., Dobreva, M. (2010). Matching evolving Hilbert spaces and language for semantic access to digital libraries. *Proceedings of12th international conference on Asia-Pacific digital libraries, ICADL 2010, Gold Coast, Australia, June 21-25, 2010.* (pp. 262-263). (LNCS 6102). Berlin: Springer. DOI: 10.1007/978-3-642-13654-2_35
  - Daránvi, S., Wittek, P. (2012). Connecting the Dots: Mass, Energy, Meaning, and Particle-Wave Duality. *Proceedings of QI-12*, *Paris June 26-29, 2012*. (pp. 207-217) (LNCS 7620). Berlin: Springer. DOI:10.1007/978-3-642-35659-9_19
  - Wittek, P., Daránvi, S., Liu, Y.H. 2014. A vector field approach to lexical semantics. *Proceedings of QI-14, Filzbach, Switzerland , June 30-July 3 2014.* (pp. 78-89) (LNCS 8951) Berlin: Springer. DOI:10.1007/978-3-319-15931-7_7
- Past work – Tool development:
  - 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
- Past work – Methodology development:
  - P. Wittek, P., Daránvi, S. (2007). Representing word semantics for IR by continuous functions. In S. Dominich and F. Kiss, editors, *Studies in Theory of Information Retrieval. Proceedings of ICTIR-07, 1st International Conference of the Theory of Information Retrieval*, pp. 149-155, October 2007.
  - Wittek, P., Tan,C.L. (2011). Compactly supported basis functions as support vector kernels for classification. *Transactions on Pattern Analysis and Machine Intelligence*, 33(10):2039 -2050. DOI: 10.1109/TPAMI.2011.28
  - Wittek, P. (2013).Two-way incremental seriation in the temporal domain with three-dimensional visualization: Making sense of evolving high-dimensional data sets. *Computational Statistics and Data Analysis*, 66:193-201.
  - Daránvi, S., Wittek,P. (2013). Demonstrating conceptual dynamics in an evolving text collection. *Journal of the American Society for Information Science and Technology*, 64(12):2564-2572. DOI: 10.1002/asi.22940
  - Wittek, P., Daránvi, S., Kontopoulos, E., Moysiadis, T., Kompatsiaris, I. (2015). Monitoring term drift based on semantic consistency in an evolving vector field. In *Proceedings of International Joint Conference on Neural Networks (IJCNN*), July 12-17, 2015. (pp 1-8.) Also at https://arxiv.org/abs/1502.01753
  - Daránvi, S., Wittek, P., Konstantinidis, K., Papadopoulos, S., Kontopoulos, E. 2016. Physics as a metaphor to study semantic drift. *Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems - SEMANTiCS2016 and the 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS'16) co-located with the 12th International Conference on Semantic Systems (SEMANTiCS 2016), Leipzig, Germany, September 12-15, 2016*. CEUR Workshop Proceedings Vol-1695, urn:nbn:de:0074-1695-3. At http://ceur-ws.org/Vol-1695/. See also at https://arxiv.org/abs/1608.01298

pericles evolving data

Roots of the metaphor

- Dynamic clustering (Salton 1975) is 2/3 of Newton's 2nd law, $F = ma$
- Similarity is *as if* glue = modelled on attractive force, e.g. gravitation, *if* term "mass" can be factored in
  - PageRank can act as term "mass"
  - PR values are variable, not constant, implementing social mechanics
- An RBF kernel 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. $K(x,x') = \exp(-\gamma||x - x'||^2)$
  - Semantic kernels hint at a deeper connection between meaning and the metric tensor shaping the curvature of classification space (Amari & Wu 1999, Eklund 2016)
  - "Lexical forces" exist (White 2002, Beeferman *et al* 1999)
- Word meaning as a lexical field: theory exists but is static
  - Dynamic semantics, update semantics is sentence semantics



$$F_1 = F_2 = G\frac{m_1 \times m_2}{r^2}$$

Dataset and tool

- Dataset: Tate Britain, London
  - 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. 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.
  - 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 6624 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 7536 unique terms (288 of them persistent over ten epochs).
  - Analysis Framework Description in the paper, see the one to the next flowchart.
- Tool: Somoclu
  - Massively parallel, fastest open source software to train self-organizing maps
  - Accelerate training on multicore CPUs, GPUs, and cluster
  - Here: Emergent Self-Organizing Maps (ESOM)

Artwork and search metadata example: J.E. Millais' *Ophelia*

Find similar objects

Artist

Sir John Everett Millais, Bt (65)

Type of object

artwork (77,273)
- painting (5,417)

Date

1851–2 (88)

Style or '-ism'

19th century (4,347)
- Pre-Raphaelite Brotherhood (547)

Subject

emotions, concepts and ideas (15,689)
- emotions and human qualities (5,133)
  - despair (209)

leisure and pastimes (6,751)
- music and entertainment (2,047)
  - singing (75)
- recreational activities (2,335)
  - picking flowers (28)

literature and fiction (3,154)
- Shakespeare (137)
  - 'Hamlet' (16)
- characters (446)
  - Ophelia (11)

nature (44,117)
- animals: birds (1,304)
  - robin (10)
- plants and flowers (2,638)
  - daisy (26)
  - forget-me-not (2)
  - fritillary (1)
  - meadowsweet (1)
  - nettle (3)
  - pansy (10)
  - pheasant's eye (1)
  - poppy (18)
  - purple loosestrife (1)
  - reed (71)
  - rose (122)
  - violet (4)
- trees (3,235)
  - willow (26)
- water: inland (11,603)
  - bank (815)
  - stream (315)

objects (22,375)
- clothing and personal effects (5,803)
  - dress (370)

people (32,377)
- actions: postures and motions (8,526)
  - floating (37)
  - hand/hands raised (140)
- adults (22,782)
  - woman (8,982)
- diseases and conditions (1,359)
  - drowned (58)
  - mental illness (178)
- named individuals (9,901)
  - Siddal, Elizabeth Eleanor (12)
- portraits (4,188)
  - individuals: female (1,592)

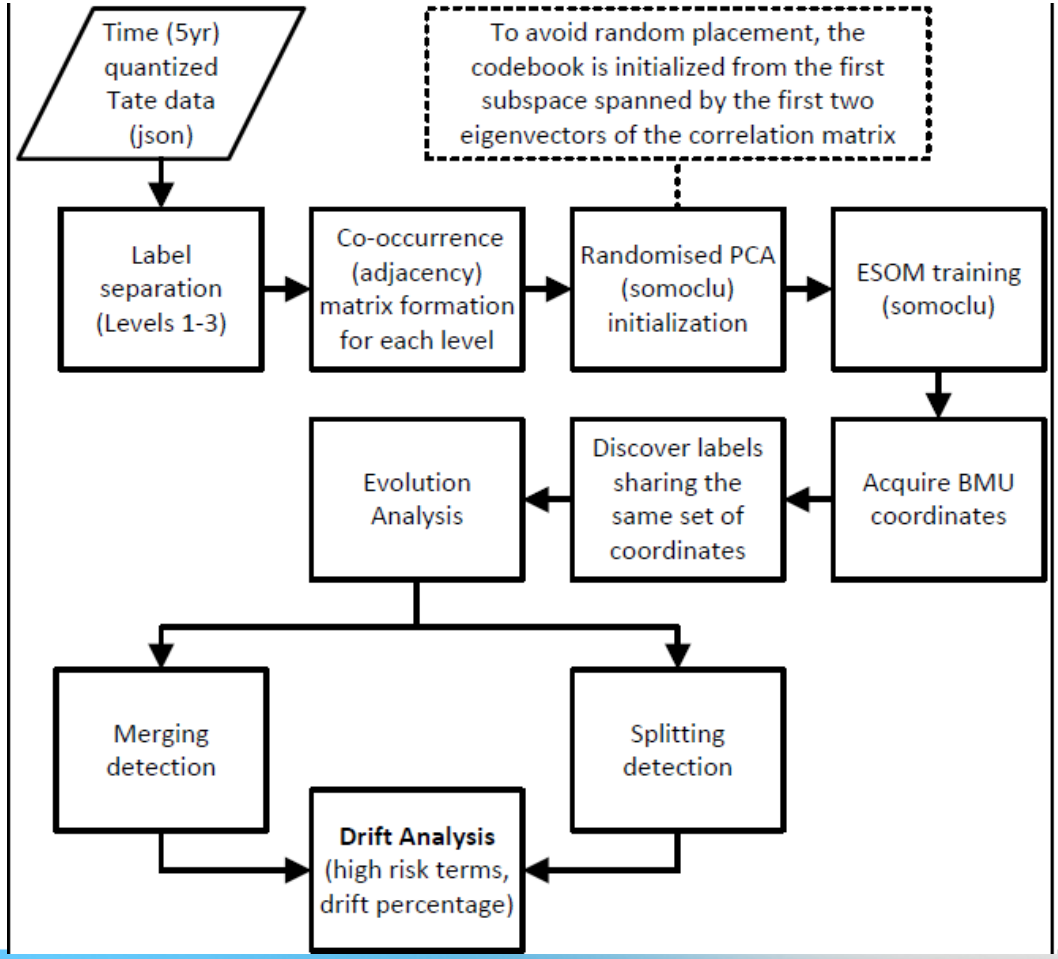
places (40,482)
- UK cities, towns and villages (12,953)
  - Ewell (3)
- UK counties (19,986)
  - Surrey (334)
- UK countries and regions (24,337)
  - England (19,083)
- UK natural features (5,672)
  - River Hogsmill (1)

- emotional states and conditions (158)
  - faithfulness - violet (1)
  - innocence - daisy (1)
  - love, foresaken - willow (1)
  - love, in vain - pansy (1)
  - memory - forget-me-not (1)
  - pain - nettle (1)
  - sorrow - fritillary (1)
  - sorrow - pheasant's eye flower (1)
  - uselessness - meadowsweet (1)

work and occupations (11,717)
- royalty and social rank (1,003)
  - courtier (72)

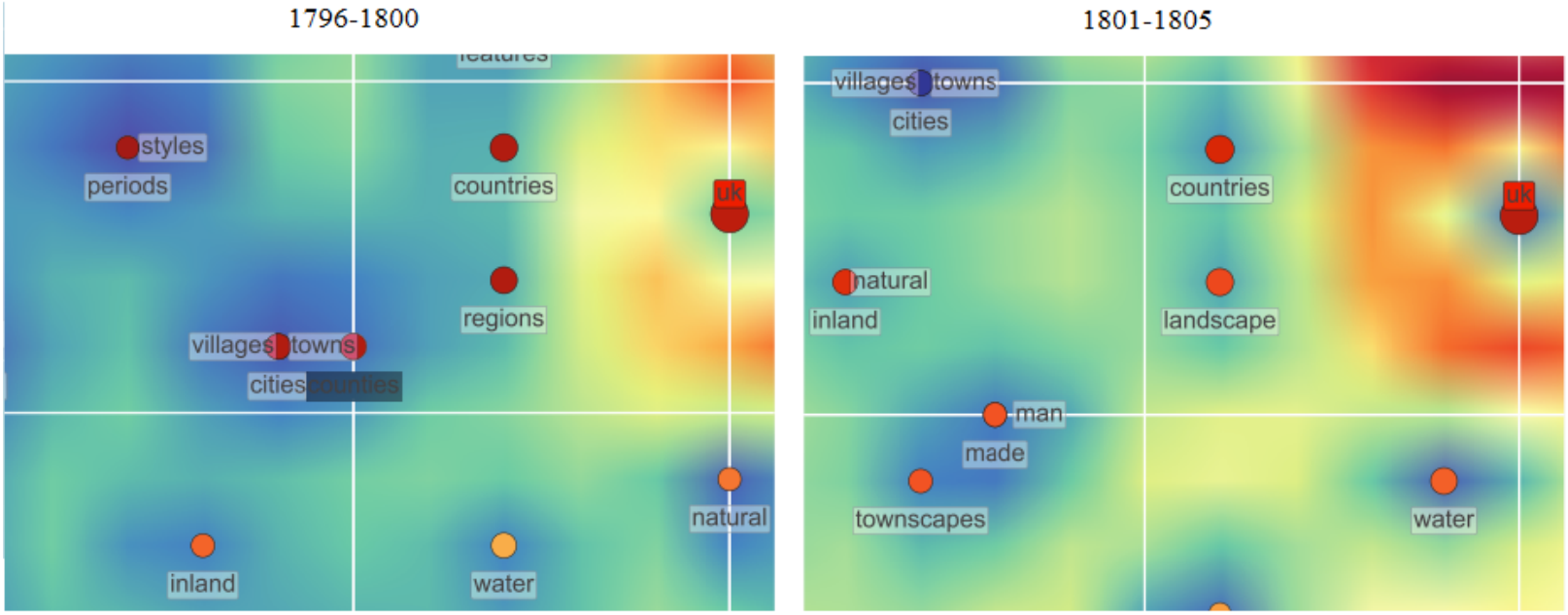
Pericles
evolving data

Tate processing flowchart



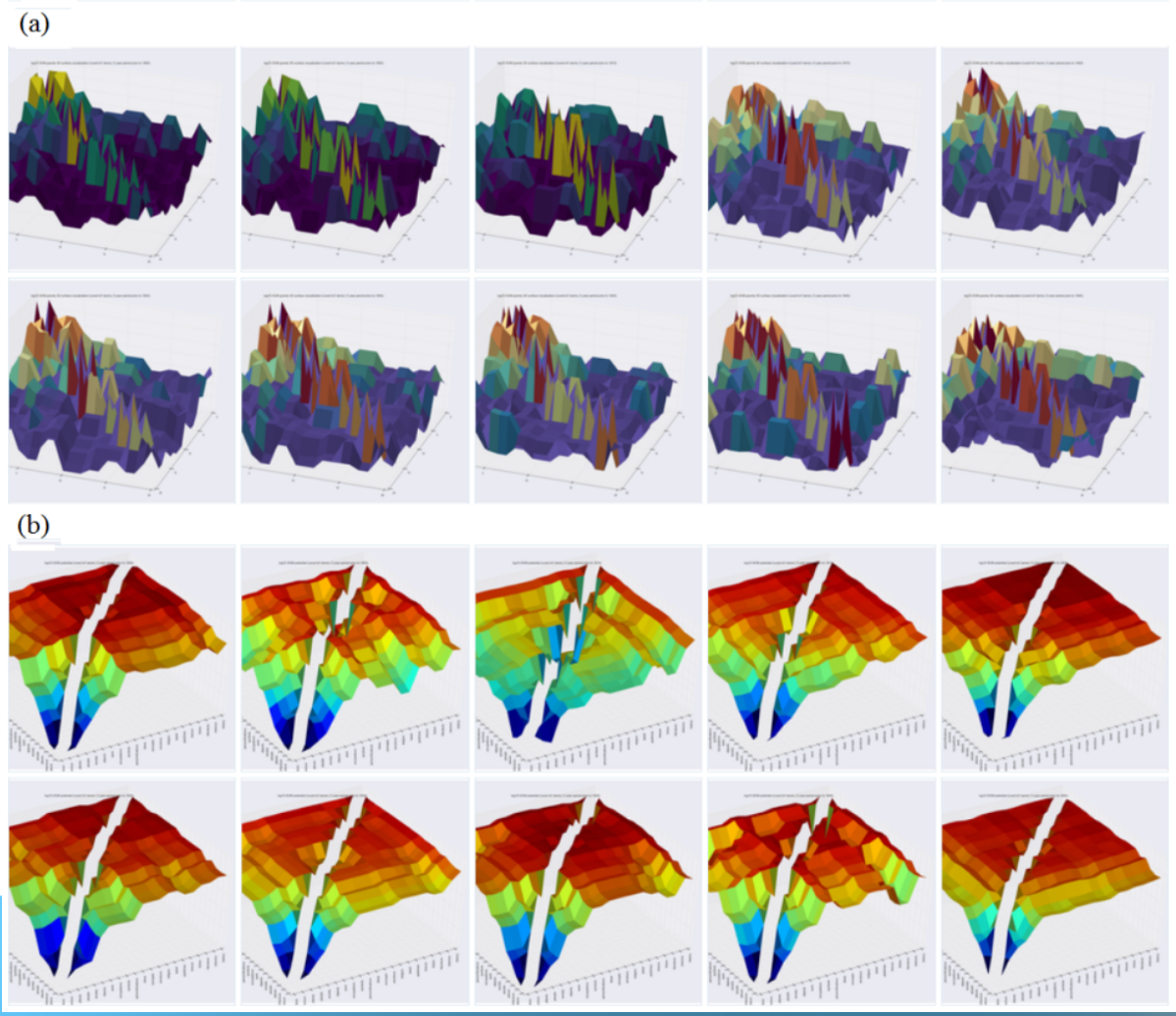
pericles
evolving data

Drifts

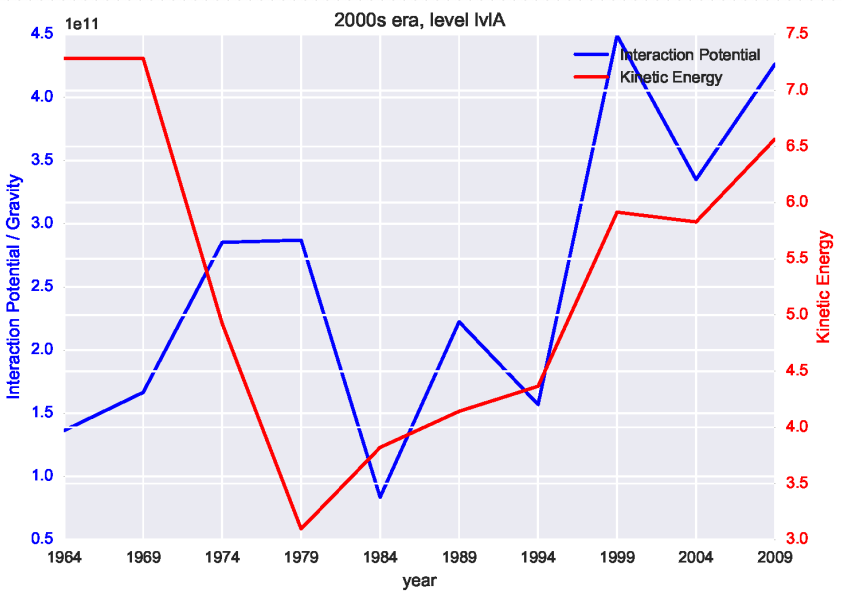


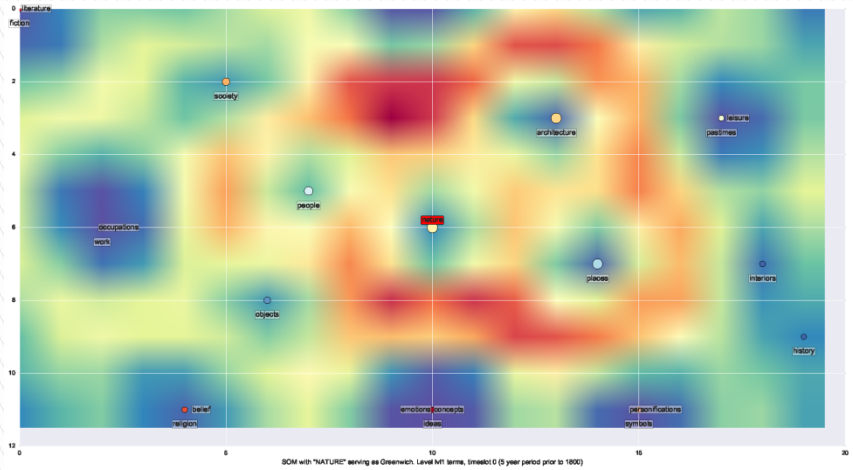
- Excerpt from the tension vs. content structure changes in the level 2 (intermediate) index term landscape in 1796–1805. Blue basins host content, brown ridges indicate tensions. Whereas *towns, cities, villages* remain merged over both epochs, *inland* and *natural* become merged within the same basin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

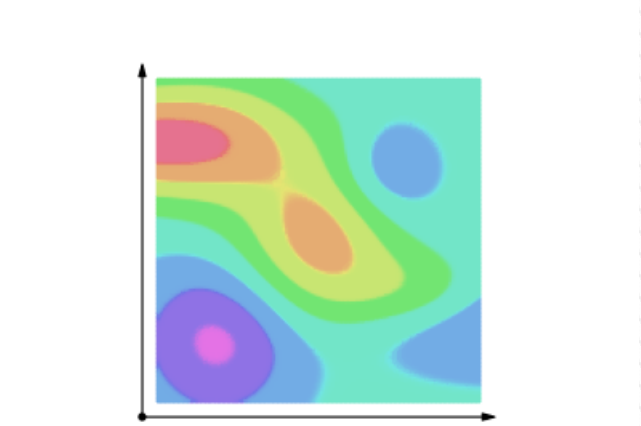
(a) Changes in the top [level 1] conceptual layer of the Tate indexing vocabulary in 1796–1845, sampled every 5 years, modelled on a gravitational field. Gravitational force is the negative gradient of the corresponding potential.
(b) Respective changes in the underlying potential field. Extreme values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



Latest outcome of work in progress







Underlying drift dynamics, we can detect external influences on timestep-specific semantic composition of collections

We can express the “work content” (i.e. energy) equivalent of converting any two meanings into each other;