Dynamics of knowledge organization

- *A 7,5 credits Doctoral course*

pericles
FP7 digital preservation

Billing Adress	Visiting Adress	Phone Number	Fax Number	E-mail
University of Borås	Allégatan 1	+46 33 435 40 00 (vxl)	+46 33 435 40 03	Nasrine.Olson@hb.se
SE-501 90 Borås		+46 33 435 4318 (direkt)		

Contents

Introduction 3
Course content 4
Implementation 5
Results 7
Addendum 1. Assessment Criteria 9
Addendum 2. Assignments 12
Addendum 3. Course Syllabus 14
Addendum 4. Reading 15
Addendum 5. Information about software Somoclu 17

Introduction

One of the results of the EU-FP7 funded project Pericles was the concept and material for a Ph.D. course developed as a part of work package 7 - Training. In this report the course and its results will be described. The course was organized by University of Borås with participation from project member CERTH in Greece.

The focus of the course is on the nature and role of knowledge organization in a rapidly changing digital world. It stresses that time series of content, prominently semantic content, have become the subject of intensive interest over the past two decades for as diverse subject areas as digital preservation, knowledge engineering, data science, natural language processing, or document engineering, to name but a few. Therefore this dynamics, something in the course called *evolving semantics*, is important for Library and Information Science as well because it influences our understanding of knowledge organization in a fundamental way.

The course approached the subject from two angles to familiarize the students with the concept of evolving semantics. Both angles focused on text-based documents written in natural language, although the model could be easily generalized to other modalities too.

One particular approach addressed was the role of ontologies as the currently ultimate level of indexing vocabularies. This direction departs from logic and formal semantics, and has led to the creation of the Semantic Web. The other approach has its background in multivariate statistics used for the automatic indexing and classification of documents, but also for information retrieval and machine learning.

When applied to language, both logic and statistics teach us about word and sentence semantics, i.e. the meaning inherent in text documents and their collections. And precisely by the tools built on and around these insights, one can detect and utilize perpetual changes that impact static knowledge structures and turn them into dynamic ones, caused by evolving semantics.

The course included core readings from both approaches to explain their fundamentally similar goals, but also to point out some key differences. Software tools developed in the PERICLES EU project were used for the students to get hands-on practice in familiarizing themselves with the subject. The course outcome for each student was a research report combining theoretical and practical skills by experimentation.

Four instructors led the course: Sándor Darányi (course leader, HB), Nasrine Olson (HB), and two guest lecturers, Efstratios Kontopoulos (CERTH/ITI, Thessaloniki, Greece), and Peter Wittek (HB, and ICFO, Barcelona, Spain). Software troubleshooting and backup was provided by Konstantinos Konstantinidis (CERTH).

Course content

The course was given at the University of Borås as a 7,5 credits doctoral course. The course has been available since autumn 2016 and is currently part of the course package for doctoral students at the Swedish School of Library and Information Science (SSLIS) at the University of Borås. Prospective students need to hold at least a bachelor´s degree and also hold the qualifications stipulated in the general study plan for doctoral studies at SSLIS to be eligible for the course.

In the course syllabus[1] is stipulated a number of learning outcomes/goals which the students are measured against in the written exams and seminars. The learning outcomes are

With respect to knowledge and understanding

- explain and account for the components of knowledge organisation affected by change, such as logical structures, indexing terminology, social context of knowledge, etc.
- demonstrate an improved understanding of similarities in, and applicability of, dynamics in deep structure in relation to surface morphologies

With respect to skills and abilities

- perform measurements in complex evolving knowledge environments
- develop new applications for accessing knowledge resources

With respect to professional judgments

- be able to assess independently and critically the strengths and limitations of a particular methodology related to evolving semantics
- be able to identify pertinent novel approaches to the problem of collection diagnostics

Contents

- Semantics and digital preservation: basic concepts, theories and trends
- Vectors and matrices: word and sentence meaning for advanced access to digital collections

These goals are complemented by an extensive reading plan which can be found in its entirety in Addendum 4. Reading. The reading is an integral part of the learning process and a foundation for the intensive tutoring period. The material ranges over these four topics.

Topic 1: Cultural and social impact of change processes
Topic 2: Ontologies and change
Topic 3: Vector field semantics
Topic 4: Conceptual clarifications

The students have access to most of the material via the study platform Ping-Pong or the university library. The required reading has a mandatory addition that the students must find complementary reading.

---

1 Addendum 3. Course Syllabus

Implementation

The course was offered on a single occasion so far and was attended by 8 students from Sweden, the UK and Croatia.

The course was conducted in a distance learning environment with one intense tutoring period held of three days at the University of Borås. During these three days students had the possibility to attend in person or via video-conference. During the entirety of the course the students and lecturers communicated by the study platform used at the University of Borås and through video-conferences held via Skype or Adobe Connect.

The course outlay is presented below.

> The intensive tutoring period will take place by videoconference on January 16-18, 2017. There will be lectures with slide presentations, and exercises for two software tools. After this intensive phase, you will have to work on a report and combine your readings with the processing of test data by both tools. This report will constitute the written home examination for the course and will have to be handed in by February 24 (Friday), 2017 midnight at latest.
>
> The course consists of two main major parts, a reading- and an experimentation-oriented one. They allow some flexibility in studies, however below we also have spread the work on different assignments in time to provide a chronological framework as a suggestion. A possible workload distribution to meet all the interim deadlines is as follows:
>
> - December 6 - 11 (week 49): Read Schlieder, selected chapters of Salton, Nöth, Lyons for a background. Take part in the course requirements discussion.
> - December 12 - 18 (w 50): Read Berners-Lee et al. (2001) and Chapter 1 from Antoniou & Van Harmelen, F. (2004).
> - December 19 - January 31 (w 51-52) plus January 1: Christmas break.
> - January 2 - 8 (w 1): Read Noy & McGuinness (2001) and sections 7.1-7.4 from Antoniou & Van Harmelen, F. (2004).
> - January 9 - 15 (w 2): Read Turney and Pantel, Ultsch and Moerchen.
> - January 16 - 22 (w 3): Intensive tutorials and training with software (Somoclu, Protege, SemaDrift).
> - January 23 - 29 (w 4): Read Stavropoulos et al. (2016), Wittek et al. (2014), Wittek et al. (2015).
> - January 30 - February 5 (w 5): Practice with Protege based on Horridge et al. (2011) and on SemaDrift and work on the home exercises.
> - February 6 - 12 (w 6): Crosscheck Osgood for a broader understanding. Research questions as the focus for the final report and practical task are published.
> - February 13 - 24 (w 7 & 8): Work on the research report. Submission deadline: February 24, 2017 (24:00).

A seminar was held at the beginning of the intensive tutorial period on the 16th of January 2017 where the students presented their previously gained knowledge regarding such subjects as the Semantic Web and its benefits, the deployment of semantic practices and their importance in knowledge environments subject to change processes.
Intensive tutoring period schedule

Jan 16 Mon Time (CET) Topic
C430 or
Online
09:00-09:30 Content dynamics and Digital Preservation (Sándor Darányi)
09:30-10:00 The Semantic Web and the emergence of ontologies (Stratos Kontopoulos)
10:15-11:00 Ontology languages and ontology engineering (Stratos Kontopoulos)
11:15-12:00 Crash course to the Protege ontology editor (Stratos Kontopoulos)
12:00-13:00 Lunch
13:00-13:45 Ontology evolution and semantic drift (Stratos Kontopoulos)
14:00-14:45 Introduction to the SemaDrift tool (Stratos Kontopoulos)
15:00-15:45 Lab exercises with SemaDrift (Stratos Kontopoulos)
16:00-17:30 Seminar (Sándor Darányi, Nasrine Olson, Stratos Kontopoulos)
Jan 17 Tue 09:00-17:00 Individual practicing and preparation for next day
Jan 18 Wed 09:00-09:45 Theories of word meaning significant for knowledge organization (Sándor Darányi)
C430 or
Online
10:00-10:45 Word meaning: from semantic fields to vector fields (Sándor Darányi)
11:00-12:00 Machine learning on unstructured data (Peter Wittek)
12:00-13:00 Lunch
13:00-13:45 Unsupervised machine learning and visualization (Peter Wittek)
14:00-14:45 Lab exercises with Somoclu 1 (Peter Wittek)
15:00-16:30 Lab exercises with Somoclu 2 (Peter Wittek)
16:30-17:00 Summing up: signs, fields and dynamics (Sándor Darányi)

Lab exercises were performed by the students on their own using the software packages Protege, SemaDrift and Somoclu after having introduced them during the intensive tutoring period. These tools are for working with the development of ontologies, and in this course with special regard to semantic drifts and their statistical analysis.

The above tools were used on catalog metadata from Tate Galleries, London, the period chosen was from their two most intense periods of acquisition, 50 years during the 19th century and 50 years during the 20th century. The ontology track of inquiry used several publicly available ontologies from the WWW.

Results

The course is a good example of how the results of a research project find their way into teaching and subsequently into the public domain. The course will continually be offered at the University of Borås, available at no cost for all EU-citizens. Respective course material from the PERICLES project and related to the course is also publicly available in the PERICLES Training Module Package and a corresponding Massive Open Online Course (MOOC) has also been tested to further disseminate the results.

Six out of the eight registered students passed the course, several of them with excellent results. The grading is based on home exams where the students integrated the theories from the reading phase of the course with the more practical exercises after the intensive tutorial period and had to report their experimental findings.

The students acquired deep knowledge of the subject to bring home to their respective organizations and countries.

The results of the course show that the content has a bearing upon the approached subject and is as such a valuable contribution to the possible knowledge base for doctoral students all over the EU.

Addendum 1. Assessment Criteria

Content

	Excellent (VG)	**Pass (G)**	**Fail (U)**
Interpretation and scope	Excellent. Good attempt to reflect scope of essay. Almost all significant points covered.	Most key points covered but some omissions and/or misunderstandings.	Inadequate attempt t define scope of essa Scope of essay misunderstood.
Understanding of topic	Excellent understanding and exposition of relevant issues. Insightful and well informed. Some awareness of nuances and complexities.	Clear awareness and exposition of relevant issues. Shows awareness of the issues but no more than to be expected from attendance at classes.	Establishes a few relevant points but superficial and conf Much irrelevant ma
Use of literature	Excellent use of evidence to support arguments/points. Some evidence of independent research.	Good use of evidence to support arguments. Over-reliance on a limited range of material.	Insufficient evidenc independent work, r on a superficial repe class notes.
Evaluation and synthesis of evidence	Excellent standard of evaluation and synthesis of source material.	Some evaluation and synthesis of source material relative to the organization.	Insufficient evaluati source material. Poo understanding of cla notes.
Critical analysis	High standard of critical analysis using appropriate conceptual framework. Questioning, unbiased approach. Clear evidence of independent thought	Uses appropriate conceptual framework. Some questioning of written sources. Attempts analysis but some omissions and/or errors.	Weak understanding conceptual framewo Too descriptive and analysis too superfic with omissions and errors.
Structure, logical development	Arguments clearly structured and logically developed.	Arguments clearly structured and logically developed.	Arguments often confused and undeveloped. No log structure.
Conclusion	Draws together various points. Identifies key issues, solutions.	Very good. Draws together main points. Conclusion does not do justice to body of essay. Too short.	No recognisable conclusion.

Presentation

	Excellent (VG)	**Pass (G)**	**Fail (U)**
Abstract	Very clear definition of subject.	Clearly defines subject.	Abstract missing or inadequate.
Spelling, grammar and syntax	Not applicable		

Style	Not applicable		
Presentation of data and references	Good selection of data and references. Relevant data and examples, all properly referenced. References accurately cited and listed.	Some good use of relevant data and examples but incompletely referenced. Occasional errors in citation missing or incorrect citations and/or bib. entries.	Superficial use of rel data and examples ar poor references. Nun of errors in citation, missing or incorrect citations and/or bib. entries.
Length	Length appropriate.	Length appropriate.	Short of the minimur limit.
Overall presentation	Excellent presentation. Carefully organised and well presented. Students' guidelines followed.	Carefully organised and well presented. Students' guidelines followed.	Unacceptable presentation. Paginat title, margins, paragr need attention.

Correspondence between Swedish and ECTS grades

Swedish grades	ECTS grades	Definition
VG “strong/solid”	A	EXCELLENT -- outstanding result with minor shortcomings/mistakes only
VG “ordinary”	B	VERY GOOD – above the average but with more shortcomings/mistakes
G “strong/solid”	C	GOOD – generally good quality work but with a number of obvious shortcomings/mistakes
G “ordinary”	D	SATISFACTORY – acceptable but wanting/with reservations
G “weak”	E	SUFFICIENT – the result meets minimal criteria
U closer to G	FX	INSUFFICIENT – more effort is necessary before credits can be assigned
U	F	INSUFFICIENT – considerably more effort is necessary before credits can be assigned

Additional Comments:

The strong points of the [essay, thesis]

The areas that could be improved

Addendum 2. Assignments

Home exercises for Somoclu on the Tate dataset

To understand the exercises below, please feel free to read again the `tate somoclu user guide and exercises_16-12-31.docx` file in the Documents/Lab/Somoclu related folder. For best results, you also have to take a look at the contents of the Tate subject index related subfolder.

Exercise 1: In the above folder, the file `1800s lvl1 drift stats_16-12-31.xlsx` contains the ranking of four parameter combinations in Somoclu to see which one reconstructs the originally compound indexing expressions (column B) from lower level index terms (column C). Add one or more of the missing parameter combinations and evaluate their results. How will the new ranking differ from the current one, listed by the end of the recovery table?

Exercise 2: Repeat the above evaluation procedure for level 1 (lvl2) terms of the 1800s acquisition series. You will find the respective template called `1800s lvl2 drift stats_16-12-31.xlsx` in the same folder.

Exercise 3 (time-consuming, therefore optional): Repeat the above evaluation procedure for level 2 (lvl2) compound index terms on the 1800s series for three GUI parameter combinations. Will the toroid hexagonal PCA combination again outperform the toroid rectangular PCA one?

<u>Seminar focus questions</u>

Please prepare a brief PowerPoint presentation (2-3 slides) according to the questions below. **During the seminar on January 16, 2017 (Monday) you will have to discuss it with your fellow students and comment on their approaches.**

A good idea is to limit both the slides and the questions and answers part to 5+5 minutes respectively.

- Your presentations should address the following questions:

  1.a In your view, what are the key benefits that the Semantic Web brings into play?

  1.b Based on a limited Internet search, please present a paradigm (e.g. a company, an online service etc.) that is already deploying Semantic Web technologies and briefly describe how these technologies are used in practice.

  2. How do you assess the importance of the semantic drift at this point in your readings?

- It should also include a discussion question (based on your readings) formulated by you and posed to your fellow students at the seminar.

Your contribution will be assessed by the teachers when it comes to grading your coursework.

Addendum 3. Course Syllabus

Swedish School of Library and Information Science
Dynamics of Knowledge Organization Course syllabus – Doctoral course
7.5 Credits
Ladok code: FBIDKO1
Version: 1
Valid from*:* Autumn 2016
Ratified by: The Committee for research education, 2016-04-28
Educational level: Research education
Research area: Library and information science (Code 50805)

Special requirements

Bachelor's Degree or the equivalent. Priority is given to doctoral students. Applicants are
required to hold the qualifications stipulated in the general study plan for doctoral studies (see
Decision 962-10-83, FoU 2010/9)

Learning outcomes

On completion of the course the students should be able to:

- ***With respect to knowledge and understanding***

● explain and account for the components of knowledge organisation affected by
change, such as logical structures, indexing terminology, social context of knowledge,
etc.

● demonstrate an improved understanding of similarities in, and applicability of,
dynamics in deep structure in relation to surface morphologies

- ***With respect to skills and abilities***

● perform measurements in complex evolving knowledge environments

● develop new applications for accessing knowledge resources

- ***With respect to professional judgments***

● be able to assess independently and critically the strengths and limitations of a
particular methodology related to evolving semantics

● be able to identify pertinent novel approaches to the problem of collection diagnostics

Contents

● Semantics and digital preservation: basic concepts, theories and trends

● Vectors and matrices: word and sentence meaning for advanced access to digital
Collections

Addendum 4. Reading

Topic 1: Cultural and social impact of change processes
Schlieder, C. (2010). Digital heritage: Semantic challenges of long-term preservation. *Semantic Web (1)1-2,*
143-147. REM: uploaded.
Topic 2: Ontologies and change
Antoniou, G., and Van Harmelen, F. (2004). *A Semantic Web Primer*. MIT press. [available here] - see
Chapter 1 for an introduction to the Semantic Web, and Chapter 7 (sections 7.1-7.4) for the basic principles
of ontology engineering.
Berners-Lee, T., Hendler, J., and Lassila, O. (2001). *The Semantic Web: A new form of Web content that is*
meaningful to computers will unleash a revolution of new possibilities. [available here]. This is the first article
introducing the Semantic Web, (co)authored by Sir Tim Berners-Lee, the very person who is considered the inventor of the World Wide Web.
Horridge, M., Knublauch, H., Rector, A., Stevens, R., and Wroe, C. (2011). *A Practical Guide To Building*
OWL Ontologies Using The Protégé-OWL Plugin and CO-ODE Tools, Edition 1.3. University of Manchester.
[available here].
Noy, N., and McGuinness, D. L. (2001). *Ontology Development 101: A Guide to Creating Your First Ontology*.
Knowledge Systems Laboratory, Stanford University. [available here].
Stavropoulos, T., Andreadis, S., Kontopoulos, E., Riga, M., Mitzias, P., Kompatsiaris, Y. (2016). *SemaDrift: A*
Protégé Plugin for Measuring Semantic Drift in Ontologies. [Slides]. Drift-a-LOD Workshop 2016, Bologna,
November 20. (In publication).
Topic 3: Vector field semantics
Turney, P.D., and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics.
Journal of Artificial Intelligence Research (37), 141-188.
Ultsch, A., and Moerchen, F. (2005). ESOM-Maps: tools for clustering, visualization, and classification with
Emergent SOM. *Technical Report Dept. of Mathematics and Computer Science.* University of Marburg,
Germany, No. 46.
Wittek, P., Darányi, S. Kontopoulos, E., Mysiadis, T., and Kompatsiaris, I. (2015). *Monitoring Term Drift*
Based on Semantic Consistency in an Evolving Vector Field. [available at: http://arxiv.org/abs/1502.01753]
Wittek, P., Darányi, S., and Liu, Y-H. (2014). A Vector Field Approach to Lexical Semantics. In: *Proceedings*
of 8th International Conference on Quantum Interaction, Filzbach, Switzerland. (June 30 - July 3, 2014).
[available at: http://link.springer.com/chapter/10.1007/978-3-319-15931-7_7]
Topic 4: Conceptual clarifications

Because the topic of the course is new and interdisciplinary, we selected a few basic books for you to brush up your preexisting knowledge on the subject. It is not compulsory but may be helpful to look into selected chapters of theirs as indicated below.

Lyons, J. (1968). *Introduction to theoretical linguistics*. New York: Cambridge University Press. (Chapter 2: The structure of language, pp. 53-98; Section 5.4: The word, pp. 194-205; Chapter 9: Semantics -- general principles, pp. 400-442; Chapter 10: Semantic structure, pp. 443-481.) *

Nöth, W. (1990). *Handbook of semiotics*. Bloomington: Indiana University Press. (2-componential theories of word meaning in Chapter II: Sign and Meaning, Section Sign, pp. 79-92.)

Osgood, C.E., Suci, G.J., and Tannenbaum, P.H. (1957). *The Measurement of Meaning*. Urbana: University of Illinois Press. (For general background only, especially Chapter 2 on semantic spaces.) **

Salton, G. (1968). *Automatic information organization and retrieval*. New York: McGraw-Hill. (Chapter 2: Information analysis and dictionary construction, pp. 21-65; Chapter 4: Statistical operations, pp. 110-148) ***

* If you cannot get hold of this book, a good replacement is Lyons, J. (1977). *Semantics* Vol 1. London: Cambridge University Press. (Chapter 1: Introduction, pp. 1-31; Chapter 6: Logical semantics, pp. 138-173; Chapter 7: Reference, sense and denotation, pp. 174-229; Chapter 8: Structural semantics I -- Semantic fields, pp. 230-269; Chapter 9: Structural semantics II -- Sense relations, pp. 270-335, all uploaded).

** For those willing to go in the statistical direction, an update on Osgood *et al*. 1957 is now Samsonovic, A.V. and Ascoli, G.A. (2010) *Principal Semantic Components of Language and the Measurement of Meaning*. PLoS ONE 5(6): e10921. doi:10.1371/journal.pone.0010921 .

*** If somebody is interested in the first attempt to combine automatic document processing with dynamics, feel free to study Salton, G. (1975). *Dynamic library and information processing*. Englewood Cliffs, N.J.: Prentice-Hall, Inc. as well.

Addendum 5. Information about software Somoclu

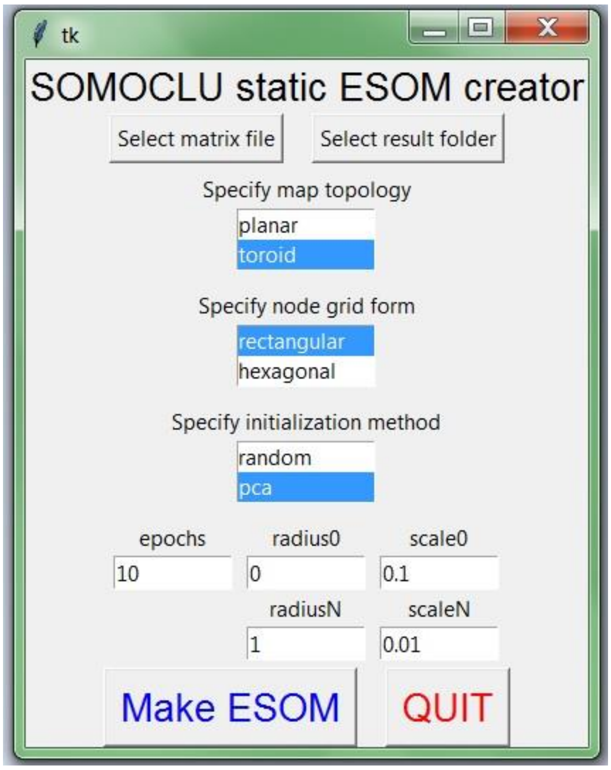
Somoclu user guide for the FBIDKO1 H16 “Dynamics of Knowledge Organization” doctoral course

Datasets: The Tate Britain catalog metadata can be found on Github. Over the past two centuries, there were two acquisition peaks, one between 1796-1845, the other between 1960-2009, the latter still being unfinished. We created two respective datasets which can be studied both for the static view of the indexing vocabulary composition at any observation point, and by a dynamic view its changes over time. Both datasets are segmented into 5-year eras. You can find them on PingPong in the Documents/Datasets/Tate dynamic co-occurrence matrices folder, both zipped and unzipped for era-specific download, or also here.

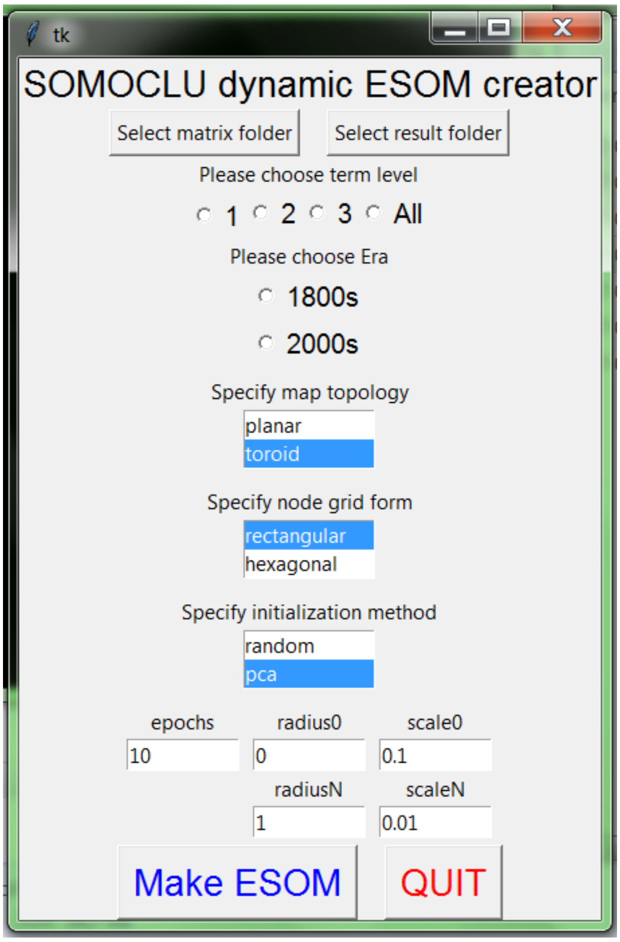
Both the 19th century and the 20th century datasets contain co-occurrences of index terms used for the appraisal of artefacts, as exemplified by any item in the collection, e.g. here, *Ophelia*. The subject index uses a shallow conceptual hierarchy of just three levels from specific over intermediary to generic concepts, indicated in our notation as *lvl1* = level 1, *lvl2* = level 2 and *lvl3* = level 3 in filenames. Thus, e.g. the filename `AdjMat2000slvl1_5.txt` refers to the top-level term co-occurrence matrix in the 5th 5-year era of the 1960-2009 period. Using this file naming convention, one can study index term drifts on any available conceptual level separately, or on every level together. In this latter case, *lvlA* in a filename such as `AdjMat1800slvlA_5.txt` indicates all conceptual levels of index terms studied in bulk.

To study these collections, you can use the Somoclu graphical user interface (GUI) tailored to Tate in both a static and a dynamic version. The static version can be launched by double-clicking on `staticSomocluWrapperGUI_vTate.py` in your working directory. The name of the respective file for the study of content dynamics is `dynamicSomocluWrapperGUI_vTate.py`, to be run in a similar fashion.

Static Tate-specific GUI variant



Dynamic Tate-specific GUI variant

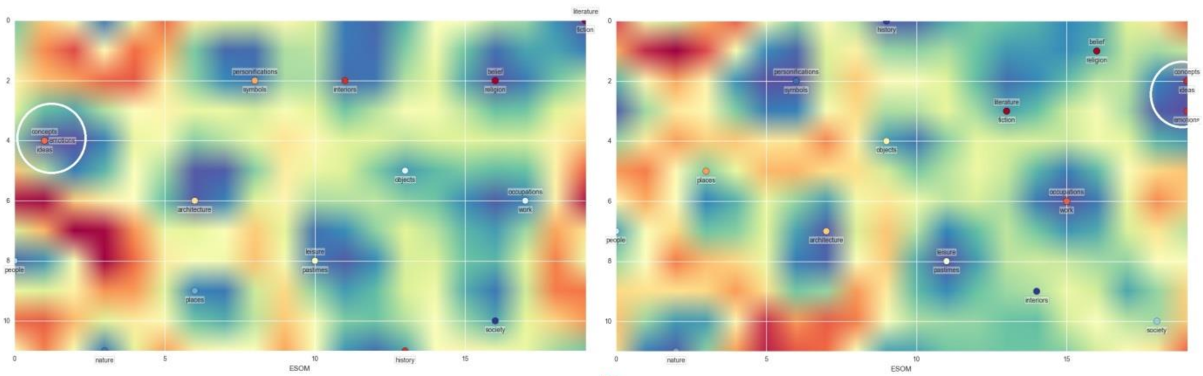


Preliminaries

Before you start experimenting, it is a good idea to create a collection-specific subfolder in your working directory where Somoclu will store the respective results. Both for static and dynamic analysis, the filenames contain the parameter combinations for easier recognition. Further, for dynamic analysis, the results also contain a drift log subfolder.

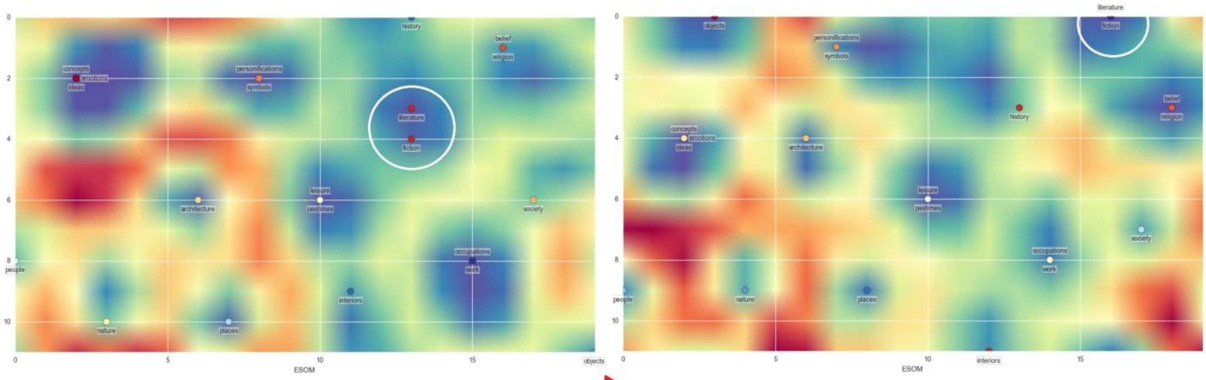
You can set the above parameters as you wish. The corresponding results will be different. However, there is also a second subtle source of difference: if you set e.g. radius to 5 and repeat the same job twice (or more), the outcomes will be slightly different. This is due to the fact that Somoclu initializes the ESOM algorithm by a random number called a *random seed*.

Semantic drifts are separated into splits and merges. *Splits* refer to a previous term label combination mapped to a single node whose one or more components became dislocated, i.e. are now mapped onto two or more different nodes. E.g. if a drift log file named `changes1800slvl1_3_1482006145.txt` states that “Terms emotions at 19,3 were split from 1,4 / Terms ideas, concepts at 19,2 were split from 1,4|1,4” this means that in the map called `1800slvl1_2_1482006145.png` the terms *emotions*, *ideas and concepts* used to be mapped on the same grid node at coordinates x =1, y =4, whereas in the next map called `1800slvl1_3_1482006145.png`, *concepts* and *ideas* are mapped now to x = 19, y = 2 while *emotions* can be found at x = 19, y = 3. This is shown below (please enlarge):



1800slvl1_2_1482006145.png → 1800slvl1_3_1482006145.png

Merges refer to the opposite, agglomerative tendency. A statement like “Terms fiction, literature at 16,0 were merged from 13,4|13,3” in the log file called `changes1800slvl1_1_1482006145.txt` indicates that two index terms which used to be located at x = 3, y = 13 vs. x = 4, y = 13 in the previous map called `1800slvl1_0_1482006145.png` became merged at the node coordinate x = 0, y = 16 in the map called `1800slvl1_1_1482006145.png`, see below (please enlarge):



1800slvl1_0_1482006145.png 1800slvl1_1_1482006145.png

Lab and home exercises with Somoclu

<u>Goal</u>: Tool calibration by drift evaluation.

<u>Means</u>: Visual inspection of drift analytic results.

<u>Reasoning and task description</u>: For every collection of composition C_t@, where t@ refers to the timestep when the observation is made, in principle or in practice there exists a corresponding ontology to define the meaning of its index terms. For the Tate catalogue, this ontology is the Tate subject index (TSI). Given an expanding collection, to update its ontology at t@+1, we inspect the respective content distribution in Somoclu maps. The research question is, how well can one recover the compound indexing expressions of TSI by statistical means? This translates to looking into grid nodes with multiple labels.

> <u>Hint</u>: The stability of SOM node content is key to vocabulary control for artefact access. Splits in and merges of node content *may* erode/recover the original meaning of the compound indexing expression. If something was indexed by diseases and conditions at t@ whereas at t@+1, diseases and conditions are now split into two separate terms, to retrieve artefacts indexed by a compound expression by its fragments becomes problematic and reduces overall content accessibility.

The file 1800s lvl1 drift stats_16-12-31.xlsx (on PingPong in the Doc/Lab/Somoclu related subfolder) is an evaluation template to rank algorithmic parameter combinations of the Tate specific dynamic GUI. By this we mean the $2^3 = 8$ available options by the respective combinations of *toroid/planar*, *rectangular/hexagonal* and *PCA/random*. As each of these 8 options recovers the original compound index terms in the TSI to a different extent, we are interested in their ranking: the one that best reconstructs the original from its text words is best suited to assist ontology maintenance.[2] To arrive at that conclusion, one chooses a multilabel node at t@ and checks it out in the next map at t@+1. If the same labels still tag a single node, we

---

[2] The TSI indexing expressions were preprocessed before statistical analysis. One of the constraints was to consider artefacts indexed by single terms, not compound ones.

enter M for *merged* in the respective cell of the evaluation table, or *S* for split otherwise. Finally we calculate the respective column and row percentages and rank the outcomes (see below).

	M = merged, S = split												
Period = 1800s			1796-1800	1801-1805	1806-1810	1811-1815	1816-1820	1821-1825	1826-1830	1831-1835	1836-1840	1841-1845	
Term level = 1	**Tate compound index terms**	**labels on nodes**	1800slvl1_0	1800slvl1_1	1800slvl1_2	1800slvl1_3	1800slvl1_4	1800slvl1_5	1800slvl1_6	1800slvl1_7	1800slvl1_8	1800slvl1_9	recovery rate (%)
GUI parameters:	symbols & personifications	[personifications, symbols]	M	M	M	M	M	M	M	M	M	M	*100*
dynamic	literature and fiction	[literature, fiction]	S	M	M	M	M	M	M	M	M	M	*90*
toroid	work and occupations	[work, occupations]	M	M	M	M	M	M	M	M	M	M	*100*
rectangular	leisure and pastimes	[leisure, pastimes]	M	M	M	M	M	M	M	M	M	M	*100*
PCA	religion and belief	[belief, religion]	M	M	M	M	M	M	M	M	M	M	*100*
	emotions, concepts and ideas	[concepts, ideas, emotions]	M	M	M	S	M	M	M	M	M	M	*90*
		recovery rate (%)	*83,333333*	*100*	*100*	*83,333333*	*100*	*100*	*100*	*100*	*100*	*100*	***96,667***

Period = 1800s			1796-1800	1801-1805	1806-1810	1811-1815	1816-1820	1821-1825	1826-1830	1831-1835	1836-1840	1841-1845	
Term level = 1	**Tate compound index terms**	**labels on nodes**	1800slvl1_0	1800slvl1_1	1800slvl1_2	1800slvl1_3	1800slvl1_4	1800slvl1_5	1800slvl1_6	1800slvl1_7	1800slvl1_8	1800slvl1_9	recovery rate (%)
GUI parameters:	symbols & personifications	[personifications, symbols]	M	M	M	M	M	M	M	M	M	M	*100*
dynamic	literature and fiction	[literature, fiction]	M	M	M	M	M	M	M	M	M	M	*100*
toroid	work and occupations	[work, occupations]	M	M	M	M	M	M	M	M	M	M	*100*
hexagonal	leisure and pastimes	[leisure, pastimes]	M	M	M	M	M	M	M	M	M	M	*100*
PCA	religion and belief	[belief, religion]	M	M	M	M	M	M	M	M	M	M	*100*
	emotions, concepts and ideas	[concepts, ideas, emotions]	M	M	M	M	M	M	M	M	M	M	*100*
		recovery rate (%)	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	*100*	***100***

Exercise 1: Add one or more of the missing parameter combinations and evaluate their results. How will the new ranking differ from the current one, listed by the end of the recovery table?

Exercise 2: Repeat the above evaluation procedure for level 1 (lvl1) terms of the 2000s acquisition series.

Exercise 3 (time-consuming, therefore optional): Repeat the above evaluation procedure for level 2 (lvl2) compound index terms on the 1800s series for three GUI parameter combinations. Will the toroid hexagonal PCA combination again outperform the toroid rectangular PCA one?

You can find an evaluation template for the 1800s series called 1800s lvl2 drift stats_16-12-31.xlsx in the Doc/Lab/Somoclu related subfolder on PingPong. If interested, alternatively you can go over to the 2000s series, but then you must start with creating a respective new template first.

Key question: How do you perceive the relationship between *conceptual dynamics* and *conceptual stability*?